Universidade de Aveiro Departamento de Matemática
2017

Alexandrina Maria da Silva

Estatística e Análise de Dados: unidade curricular do ensino superior em Timor Lorosa'e

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica da Doutora Sónia Gouveia, Professora Auxiliar Convidada do Departamento de Matemática da Universidade de Aveiro

Dedico este trabalho ao meu amado marido, aos meus amados filhos e aos meus amados pais pelo incansável apoio.

o júri

presidente

Prof. Doutor Pedro Filipe Pessoa Macedo

professor auxiliar do Departamento Matemática da Universidade de Aveiro

Profa. Doutora Magda Sofia Valéiro Monteiro

professora adjunta da Escola Superior de Tecnologia e Gestão de Águeda da Universidade de Aveiro

Profa. Doutora Sónia Cristina Alexandre Gouveia

professora auxiliar Convidada do Departamento Matemática da Universidade de Aveiro

agradecimentos

A Deus, pelo maravilhoso dom da vida.

À Universidade de Aveiro (UA) e, em particular, ao Departamento de Matemática pela oportunidade e pelas excelentes condições proporcionadas para realização do curso de mestrado em Matemática e Aplicações.

À Universidade Nacional de Timor Lorosa'e (UNTL) pela concessão de bolsa de mestrado e, na parceira com a Universidade de Aveiro, pelo sentido de cooperação continuar para capacitar os docentes.

À minha família, pelo incansável apoio, principalmente ao meu marido e aos meus filhos e aos meus pais. Obrigada pelo incentivo, amor e compreensão nesta etapa importante da minha vida.

À minha orientadora, Professora Doutora Sónia Cristina Alexandre Gouveia, agradeço o seu trabalho dedicação, paciência, disponibilidade, sabedoria, atenção, incentivo e, acima de tudo, as suas palavras encorajamento e estímulo. Muito obrigada por tudo.

Aos professores do Departamento de Matemática e todos os professores das disciplinas de Mestrado por partilham os seus conhecimentos científicos pelos apoios, envolvimentos, disponibilidades, incentivos, otimismos e encorajamentos demonstrados.

À Professora Doutora Clara Maria Magalhães, Doutor Ângelo Ferreira, Engenheiro Miguel de Oliveira, Professor Doutor Agostinho Miguel Mendes Agra, Professor Doutor Domingos Cardoso pelos seus máximos apoios. Meus caros amigos conterrâneas e compatriotas timorenses aqui em Portugal especialmente neste Universidade de Aveiro.

Enfim, a todos que estiveram presentes direta ou indiretamente nesta fase de minha vida e que contribuíram para a realização deste trabalho.

Muito obrigada!

palavras-chave Análise exploratória de dados, análise de tabelas de contingências, regressão linear e análise de variância.

resumo

Como novo país, Timor Leste necessita de recursos humanos adequados para o desenvolvimento sustentável do país. Desde a sua independência, em 2002, nenhuma das universidades existentes possuía uma unidade orgânica disseminadora de Ciências Exatas. A Universidade Nacional de Timor Lorosa'e (UNTL) é uma instituição do Ensino Superior Público e única no país a ter uma nova unidade orgânica, cuja visão e missão estão centradas na disseminação e promoção de Ciências Exatas no ensino superior timorense. Esta nova unidade orgânica é designada por Faculdade de Ciências Exatas (FCE) e implementa, desde 2015, o curso de Licenciatura em Ciências Exatas (LCE) com habilitação em Matemática, Física e Química. Para criar a FCE e implementar a LCE, a UNTL contou e conta com o apoio técnico e pedagógico da UA. Em particular, a UA tem recebido docentes da UNTL para realização de estudos de pós-graduação em áreas relevantes para a nova faculdade, proporcionando-lhes a atualização de conhecimentos e o desenvolvimento de competências essenciais para uma implementação de sucesso de cada unidade curricular (UC) da LCE.

Nesta tese, é apresentado o trabalho desenvolvido na preparação da UC de "***Estatística e Análise de Dados (EAD)***". A UC de EAD consiste numa disciplina avançada de Estatística com 8 ECTS, lecionada no IX semestre de LCE (habilitação em Matemática) e cuja primeira edição irá realizar-se no ano letivo 2019. O planeamento geral e os conteúdos desta UC foram desenvolvidos de acordo com o Plano de estudo definido pela FCE e, neste trabalho, foi dado especial ênfase ao estudo aprofundado das matérias a lecionar e ao desenvolvimento de material pedagógico para exposição teórico-prática e para treino prático (em contexto de aula e para estudo autónomo). Em particular, esta tese apresenta o *texto de apoio* e o *material de apoio na sala de aula* (slides de exposição, folhas de exercícios práticos e atividades para aprendizagem/treino com R) que servirão como guião para o docente e para os alunos.

A disciplina de EAD foi desenhada em 3 capítulos temáticos de competências.

O primeiro capítulo, apresenta técnicas de análise exploratória de dados e resumo de informação, para desenvolvimento de literacia na leitura dos gráficos, e na aptidão para executar análises descritivas e visualização. O segundo e terceiro capítulos dão relevância à quantificação do grau de relacionamento entre varáveis e à construção de modelos estatísticos que representem adequadamente relações entre variáveis. Em particular, estudam-se modelo log-lineares (para variáveis qualitativas), modelos de regressão linear (para variáveis quantitativas) e análise de variância (ANOVA para uma variável dependente quantitativa e uma variável independente quantitativa).

keywords Exploratory data analysis, analysis of contingency tables, linear regression, analysis of variance.

abstract As a new country, East Timor needs adequate human resources for the sustainable development of the country. Since its independence in 2002, none of the existing universities had an organic unit disseminating Exact Sciences. The National University of Timor Lorosa'e (UNTL) is an institution of Public Higher Education and unique in the country to have a new organic unit, after signing a cooperation protocol with the University of Aveiro (UA) on January 2014, whose vision and mission are centered in the promotion of Exact Sciences in Timorese higher education. This new organic unit is designated by Faculdade de Ciências Exatas (FCE) and has implemented, since 2015, the Graduation in Exact Sciences (LCE) course with qualification in Mathematics, Physics and Chemistry. In order to create the FCE and implement the LCE, the UNTL has the technical and educational support of the UA. In particular, the UA has received UNTL professors to carry out postgraduate studies in areas relevant to the new faculty, thus providing them with a knowledge update and the development of the skills for a successful implementation of each curricular unit (UC) of the LCE.

In this thesis, the work developed for the preparation of the UC of "Estatística e Análise de Dados (EAD)" is presented. The EAD consists of an advanced Statistics course with 8 ECTS, in the IX semester of LCE (qualification in Mathematics) and its first edition will take place in 2019. The general planning and contents of this unit were developed according to the study plan defined by the Faculty and, in this work, special emphasis was given to the knowledge improvement of the subjects and the development of pedagogical material for theoretical-practical exposition and for training (in class and for autonomous study). In particular, supporting text and classroom support materials (presentation slides, hands-on practice sheets and R-learning / training activities) will be presented as a guide for teachers and students.

The EAD course was designed in 3 thematic chapters of competencies.

The first chapter presents techniques for exploratory data analysis and summarize information in order to develop literacy in reading the graphs and increase the ability to perform descriptive analyzes and visualization. The second and third chapters give relevance to the quantification of the degree of relationship between variables and to the construction of statistical models that adequately represent relations between variables. Special attention is given to log-linear models (for qualitative variables), linear regression models (for quantitative variables) and analysis of variance (ANOVA for a quantitative dependent variable and an independent quantitative variable).

Índice

Lista de Tabelas

Lista de Gráficos

Parte I Introdução

1. Apresentação da Unidade Curricular (U.C.) de Estatística e Análise de Dados (EAD)

2. Objetivos e metodologias usadas neste trabalho

3. Estrutura desta tese

Parte I Introdução

Timor Leste é um novo país que restaurou a sua independência em 2002. Como o novo país, Timor Leste necessita de recursos humanos adequados para o desenvolvimento sustentável do país. Timor Leste tem os recursos naturais como o petróleo, gás natural, minerais e material lignocelulósico. Assim, é necessária uma formação fundamental de recursos humanos para transformar os recursos naturais. A área de ciências exatas básicas como matemática, física e química tornam-se crucial neste âmbito desde que os processos transformativos de matérias primas (i.e., de recursos naturais) ocorram em indústrias (químicas, físicas ou biológicas). A Ciência Exata trabalha baseando-se em cálculos, fórmulas e hipóteses para chegar a resultados precisos, por meio de métodos rigorosos que proporcionam a precisão. Os conhecimentos de ciências exatas são primordiais não só para entender o funcionamento microscópica dos processos envolvidos, mas também para avaliar e/ou modificar o desempenho das tecnologias utilizadas. Assim pode-se dizer que Ciências Exatas são a espinha dorsal do desenvolvimento de um país.

Desde inicio da sua independência, Timor conta com várias instituições de ensino superior privado e uma universidade pública, que é Universidade Nacional de Timor Lorosa'e (UNTL, http://untl.edu.tl/pt/), mas nenhuma das universidades privadas possui uma unidade orgânica disseminadora de ciências exatas. A UNTL foi uma universidade que pela primeira vez na sua história como uma instituição do Ensino Superior Pública e única no país a ter uma nova unidade orgânica, após a assinatura do protocolo de cooperação entre Universidade de Aveiro (UA, http://www.ua.pt/) e UNTL em Janeiro de 2014, cuja visão e missão estão centradas na disseminação e promoção de Ciências Exatas no ensino superior timorense. Esta nova unidade orgânica será a criação da faculdade de Ciências Exatas (FCE, http://untl.edu.tl/pt/ensino/faculdades/ciencias-exatas) e a implementação do respetivo curso inaugural de Licenciatura em Ciências Exatas (LCE) com habilitação em Matemática, Física e Química. Este curso de LCE organiza-se em torno do fenómeno das ciências exatas básicas e os formalismos que o fundamentam, lançando assim as bases para uma abordagem rigorosa e produtiva ao desenvolvimento de ciência e tecnologia. As três áreas científicas são estruturantes na aquisição de competências para recolher e interpretar informações científicas relevantes, compreender e desenhar modelos de explicações do mundo real, e ainda na capacitação para produzir julgamentos a partir de uma reflexão baseada em aspetos científicos e éticos.

Como referido no Dossiê de Curso, uma formação estruturante em matemática é essencial não só para a aquisição das capacidades de abstração, dedução e formalização, mas também para o desenvolvimento de técnicas de resolução de problemas modelados matematicamente em contextos teóricos apropriados. Nesse sentido, é necessário fornecer uma cultura matemática sólida e abrangente, fundamental e aplicada, complementada com o treino eficiente para a utilização de recursos de computação atuais.

Com a criação da FCE e a implementação do respetivo curso, a UNTL torna-se o disseminador pioneiro de ciências exatas em Timor Leste, podendo proporcionar no final do curso soluções exequíveis para mitigar a escassez de qualificações em ciências exatas necessárias ao impulso da materialização dos pilares de desenvolvimento do Plano Estratégico de Desenvolvimento (PED), concretamente da expansão eficaz dos programas nacionais em áreas industriais e de investigação científica.

Para criar a FCE, a UNTL conta com o apoio técnico e pedagógico da UA. No processo de implementação da FCE, a UA é responsável pela identificação das infraestruturas e dos equipamentos necessários à criação desta faculdade, pela elaboração do currículo desse curso inaugural em parceria com professores timorenses, pela lecionação de parte das unidades curriculares nos primeiros anos, e também pela formação, numa primeira fase na UA, de um corpo docente da faculdade.

Dando seguimento às orientações do Currículo Padrão Mínimo, da Direção Geral do Ensino Superior e da Direção Nacional do Currículo do Ensino Superior, como referido no Dossiê de Curso, as unidades curriculares são distribuídas em quatro tipos de matérias de formação: institucionais; de base; profissionais; e de especialização.

No âmbito das disciplinas institucionais, a UNTL oferece um conjunto de unidades curriculares transversais no primeiro ano curricular dos cursos, cujo objetivo é desenvolver competências nas línguas portuguesa e tétum, pensamento lógico e crítico e valores cívicos. As unidades curriculares de base, contêm os conteúdos que dão conhecimento e compreensão teóricas básicos para tornar a seu cargo materiais profissionais. As unidades curriculares profissionais contêm as matérias que caracterizam a identidade do curso que integra áreas de conhecimentos relacionadas com uma particular profissão. As unidades curriculares de especialização contêm conteúdos específicos para aumentar a compreensão e a competência em desempenhar uma matérias especificas.

Ao nível da organização curricular, a LCE terá a duração de 10 semestres (5 anos), correspondentes à obtenção de 300 créditos ECTS (Matemática: 67, Física: 65, Química: 66, Informática: 16, Língua Portuguesa: 16, Língua Tétum: 4, Língua Inglesa: 16, Direito: 4, Projeto: 14, Menores: 16, Especialização: 16). Os 16 créditos em especialização estão associados à habilitação pretendida pelo estudante e, em particular, para a habilitação em Matemática é obrigatória a frequência às UC de

- Modelação Matemática e Otimização (8 ECTS)
- Estatística e Análise de Dados (8 ECTS)

ambas a decorrer no semestre IX da LCE.

Nesta tese, é apresentado o trabalho desenvolvido na preparação da UC **de "*Estatística e Análise de dados (EAD)*"**, cuja primeira edição irá realizar-se no ano letivo 2019. Uma das condições ou requisitos para entender melhor nesta disciplina, é ter conhecimento básico sobre probabilidades. Este requisito é lecionado numa disciplina das unidades curriculares

profissionais, nomeadamente Probabilidade e Estatística, que corresponde a 8 ECTS dos 67 ECTS na área do conhecimento de Matemática.

A Estatística é considerada um ramo da matemática, que tem como principais objetivos obter, organizar e analisar dados, determinar as correlações entre eles, proporcionando conclusões e previsões. É também uma ciência de desenvolvimento de conhecimento humano através do uso de dados empíricos. Baseia-se na teoria estatística, um ramo da matemática aplicada. Na teoria estatística, o aleatório e a incerteza são modelados pela teoria das probabilidades. Algumas práticas estatísticas incluem, por exemplo, o planeamento, o resumo de informação recolhida e a interpretação de resultados. Assim, a Estatística é uma ferramenta essencial para qualquer profissional que necessite analisar informação. Por este motivo, a existência de Unidades Curriculares de Estatística e Análise de Dados em oferta formativa no ensino superior é incontornável.("Matemática elementar/Estatística - Wikilivros," n.d.)

A análise estatística desenvolvida no contexto desse trabalho foi baseada do Plano de Estudo definido pela faculdade.

1. Apresentação da UC de Estatística e Análise de Dados (EAD)

Nesta secção apresentam-se as características da Unidade Curricular de Estatística e Análise de Dados segundo os elementos que constam no Dossiê de Curso da LCE, incluindo objetivos, conteúdos programáticos, metodologias de ensino, avaliação, recursos (laboratoriais e equipamentos) e bibliografia.

FICHA DAS UNIDADES CURRICULARES DE MATEMÁTICA

Unidade Curricular	**Estatística e Análise de Dados**				Área científica/ código	Matemática	
Curso de	Licenciatura em Ciências Exatas				Departamento	Ciências Exatas	
Ano letivo	2019		Semestre		ímpar	Ano curricular	5º
Créditos ECTS	8		Tipo		Especialização		
		AT/ATP	02h00	AP	04h00	EA	04h45
Horas totais de contactos (15 semanas)	90h		Horas totais de trabalho (contacto + EA)				156h45
AT = Aula teórica; AP = Aula prática; ATP = Aula Teórica Prática; EA = Estudo Autónomo							
Nome do docente	**Alexandrina Maria da Silva**						

OBJETIVOS DA UC / COMPETÊNCIAS A DESENVOLVER
O principal objetivo é dar uma formação básica sobre análise exploratória de dados, análise de dados categóricos e modelação estatística.

CONTEÚDOS ESPECÍFICOS DA UNIDADE CURRICULAR

1. **Análise exploratória de dados**
   - i) Revisão de conceitos de estatística descritiva.
   - ii) Organização de dados: distribuições empíricas de frequências e suas representações tabelar e gráficas.
   - iii) Representação gráficas para dados classificados.
2. **Análise de tabelas de contingência**
   - i) Teste de independência de Qui-quadrado e teste de razão verossimilhança.
   - ii) Tabelas 2 x 2 – teste exato de Fisher e teste de McNemar. Medidas de associação.
   - iii) Tabelas r x c – casos particulares, localização de fontes de dependência e análise de resíduos.
   - iv) Modelos log-lineares para tabelas de contingência: caso bidimensional e caso tridimensional.
3. **Regressão e análise de variâncias (ANOVA)**
   - i) Regressão linear simples.
     - (1) Introdução.
     - (2) Estimação e inferência sobre os parâmetros.
     - (3) Avaliação da qualidade e significado da regressão.
     - (4) Validação dos pressupostos da regressão.
     - (5) Previsão.
   - ii) Regressão linear múltipla.
     - (1) Introdução.
     - (2) Estimação e inferência sobre os parâmetros.
     - (3) Avaliação da qualidade e significado da regressão.
     - (4) Validação dos pressupostos da regressão.
     - (5) Previsão.
   - iii) Análise de variância.
     - (1) Introdução.
     - (2) Análise de variância com um fator e efeito fixos.
     - (3) Análise de variância com um fator e efeito aleatórios.
     - (4) Análise de variância não paramétrica.

METODOLOGIA DE ENSINO

Aulas teórico-práticas, em que os sucessivos tópicos programáticos são apresentados pela docente e ilustrados através de exemplos e da resolução de exercícios propostos aos alunos. Sempre que for adequado as aulas serão de natureza laboratorial com utilização do software de estatística – por exemplo o SPSS ou o R.

<table>
<tr><th>AVALIAÇÃO CONTÍNUA</th><th>AVALIAÇÃO POR EXAME FINAL</th><th>AVALIAÇÃO POR EXAME EM ÉPOCA DE RECURSO</th></tr>
<tr>
<td>O estudante que pretende ser avaliado em regime de avaliação contínua deve estar presente em mais de 75% do total de aulas realizadas.
O estudante que não preencher essa quota, não tem direito de seguir a avaliação contínua.
<table>
<tr><th>Critério de avaliação*</th><th>Valor %</th></tr>
<tr><td></td><td></td></tr>
<tr><td></td><td></td></tr>
<tr><td></td><td></td></tr>
</table>
*A definir no quinto ano pelo regente de UC
O estudante que falte a uma das provas não tem direito de seguir a avaliação contínua.</td>
<td>Esta avaliação aplica-se ao estudante que faltou 25% das aulas realizadas ou que faltou a uma das provas de avaliação contínua.</td>
<td>A avaliação por exame em época de recurso destina-se aos alunos que não ficaram aprovados no regime de Avaliação Contínua.
O estudante que teve uma classificação quantitativa entre 0,0 e 3,9 valores deve realizar a Prova de Recurso.
Também se destina aos alunos que não ficaram aprovados no regime de Avaliação Final.</td>
</tr>
<tr><td colspan="3">Resultado final das classificações em avaliação contínua / Exame Final / Prova de Recurso:
<table>
<tr><th>Classificação quantitativa</th><th>Resultado</th></tr>
<tr><td>8,5 – 10</td><td>Aprovado</td></tr>
<tr><td>7,0 – 8,4</td><td>Aprovado</td></tr>
<tr><td>5,5 – 6,9</td><td>Aprovado</td></tr>
<tr><td>4,0 – 5,4</td><td>Prova de recurso</td></tr>
<tr><td>0 – 3,9</td><td>Prova de recurso</td></tr>
</table>
</td></tr>
<tr><td colspan="3">Ao estudante que comparece a todas as provas do regime de avaliação contínua é vedada a possibilidade de contabilizar o exame final como elemento de avaliação único para determinar a classificação final.</td></tr>
<tr><td colspan="3">O recurso a um regime de avaliação baseado apenas em exame final pressupõe o não aproveitamento, por falta de comparência, a pelo menos uma das provas de avaliação contínua.</td></tr>
</table>

RECURSOS (LABORATORIAIS E DE EQUIPAMENTO)
Laboratório de Informática com software SPSS e R instalados nos computadores

BIBLIOGRAFIA
1. Everitt, RB. e Torsten, H., 2010. *A Handbook of Statistical Analysis Using R*. CRC Press. 2. Fahrmeir, L., Kneib, T., Lang, S. e Marx, B., 2013. *Regression: Models, Methods and Applications*. Springer-Verlag, Berlin Heidelberg. 3. Hall, A. Neves, C. e Pereira, A., 2011. *Grande Maratona de Estatística no SPSS*. Escolar Editora, Lisboa. 4. Murteira, B., Ribeiro, C.S., Silva, J.A. e Pimenta, C., 2010. *Introdução à Estatística*. McGraw-Hill de Portugal. Lisboa.

A ordem dos conteúdos que se irão desenvolver no texto de apoio apresentado neste trabalho tem pequenas diferenças com os conteúdos programáticos no dossiê de LCE, tendo em conta com o seguimento dos conteúdos que irão ser lecionados.

Como apresentado anteriormente, os conteúdos específicos desta Unidade Curricular estão agrupados em 3 tópicos: 1) Análise exploratória de dados, 2) Análise de tabelas de contingência e 3) Regressão e análise de variância (ANOVA). O tópico de "Análise exploratória de dados" apresenta técnicas de análise exploratória de dados (qualitativos e quantitativos) e métodos para resumo gráfico de informação com objetivo de desenvolver aptidão para executar análises estatísticas. O tópico de "Análise de tabelas de contingência" apresenta técnicas para analisar e quantificar o grau de relacionamento entre variáveis (qualitativas), e métodos para construção de modelos estatísticos que traduzam a dependência entre essas variáveis. Finalmente, o tópico de "Regressão e análise de variâncias (ANOVA)" apresenta o modelo de regressão linear para traduzir a relação entre variáveis quantitativas e técnicas ANOVA para analisar o relacionamento entre variáveis independentes qualitativas e variáveis dependentes quantitativas.

2. Objetivos e metodologias usadas neste trabalho

Tendo em conta que a FCE é uma nova faculdade, que ainda não possui facilidades suficientes para apoiar no processo de ensino e aprendizagem, incluindo os materiais pedagógicos e as referências para cada Unidade Curricular (UC). Então, pretende-se este trabalho com o objetivo a desenhar uma UC, de "*Estatística e Análise de dados (EAD)*", para lecionação na Faculdade de Ciências Exatas, Universidade Nacional de Timor Lorosa'e (FCE-UNTL). Será dado especial ênfase ao estudo aprofundado das matérias a lecionar e ao desenvolvimento de material pedagógico para exposição teórico-prática e para treino prático (em contexto de aula e para estudo autónomo).

Este UC de EAD é constituído por um "texto de apoio" e vários "materiais de apoios na sala de aula" (incluindo slides, folhas com exercícios práticos e atividades para aprendizagem com R) para responder às necessidades tanto dos docentes como dos alunos. Estes materiais pedagógicos servirão como guião para os docentes e ao mesmo tempo, permitirá aos alunos em entender melhor a disciplina em causa. Além disso, parte deste trabalho também servirá como apontamentos

preparados para responder às necessidades dos alunos de disciplina de EAD, procurando incentivar os alunos a usar ferramentas computacionais, como folhas de cálculo (Excel) e o software livre e gratuito R**,** para a resolução dos vários problemas práticos. O material pedagógico foi desenvolvido em português e constitui uma grande contribuição para o curso de LCE e para a unidade curricular de EAD, tendo em conta que ainda não existia um manual de apoio em português para esta UC.

A metodologia utilizada neste projeto será predominantemente a revisão de literatura, sobretudo da análise de documentos e dos materiais didáticos para elaborar o texto de apoio e o slide de apoio. A partir desta leitura, preparar-se-á e desenvolver-se-á o texto e slide de apoio de modo a que esses possam ser utilizado no processo de aprendizagem na Faculdade da Ciências Exatas da UNTL.

3. Estrutura desta tese

Esta tese está organizada em quatro partes.

A Parte I é constituída por uma introdução, a apresentação da UC de EAD, a descrição dos objetivos e das metodologias usadas neste trabalho e, finaliza, com a estrutura desta tese.

A Parte II apresenta o texto de apoio de EAD. Nesta parte, a disciplina de EAD foi organizada segundo o Plano de Estudo definido pela faculdade, e desenhada em 3 capítulos temáticos de competências. No capítulo I, desenvolvem-se sobre análise exploratória de dados: encerra uma séria de técnicas utilizadas na análise preliminar de dados estatísticos. Basicamente apresentam-se medidas de localização, dispersão e simetria bem como técnicas gráficas de ampla utilização. No capítulo II, análise de tabelas de contingência, desenvolvem-se sobre: teste de Qui-quadrado, teste de razão verossimilhanças e medidas de associação, teste exato de Fisher e teste de McNemar, localização de fontes de dependência e análise de resíduos, modelos log-lineares para tabelas de contingência: caso bidimensional e tridimensional. No capítulo III, desenvolvem-se sobre regressão linear e análise de variância, que incluindo: regressão linear simples, regressão linear múltipla e análise de variância (ANOVA).

A Parte III apresenta os materiais de apoio na sala de aulas, como os slides de apoio para cada capítulo (de acordo com os capítulos na parte II). Em todo o conteúdo, exemplos resolvidos e explicados vão surgido à medida que os diversos assuntos vão sendo apresentados, terminado cada capítulo com exercícios, folhas práticas e atividades para aprendizagem com R.

A Parte IV apresenta a conclusão final deste trabalho.

Parte II Texto de Apoio de EAD

Introdução

O texto de apoio de “Estatística e Análise de Dados” (EAD) encontra-se divido em 3 capítulos temáticos de competências.

O primeiro capítulo, apresenta técnicas de análise exploratória de dados (qualitativos e quantitativos). Neste capítulo desenvolvem-se em detalhe diferentes modos de apresentação de dados, atráves de tabelas de distribuição de frequências e gráficos, como elaborar gráficos para resumo visual da informação contida num conjunto de dados, como obter medidas descritivas num conjunto de dados, resumindo num só número algumas propriedades do conjunto (como a localização central – por exemplo, uma média – ou uma medida de dispersão – por exemplo, um desvio padrão).

O segundo capítulo apresenta análises sobre tabelas de contingência. Desenvolvem-se em detalhe testes de hipóteses de independência entre variáveis qualitativas (teste de qui quadrado, teste de razão verosimilhanças e testes alternativos do teste de qui quadrado), como quantificar o grau de associação entre estas variáveis, e ainda análise de resíduos com objectivo para obter quais são as células que mais contribuem para a dependência. Finalmente, apresentam-se modelos log-lineares para tabelas de contingência, que permitem averiguar a existência, ou não, de dependência entre variáveis qualitativas e quantificar os efeitos das suas categorias.

O terceiro capítulo apresenta modelos de regressão linear e métodos para análise de variância (ANOVA). Neste capítulo desenvolve-se em detalhe sobre o modelo de regressão linear simples e múltiplo (para variáveis quantitativas), incluindo a definição do modelo, estimação e inferência dos seus parâmetros, significado e avaliação da qualidade da regressão, validação de pressupostos da análise e previsão de valores. Na regressão linear múltipla, ainda se inclui o diagnóstico de ponto de influentes, colinearidade entre variáveis e seleção de variáveis. Finalmente, a ANOVA (variáveis independentes qualitativas e variáveis dependentes quantitativas) trabalha-se para amostras independentes, mais especificamente ANOVA com um fator (de efeitos fixos ou de efeitos aleatórios) e ANOVA não paramétrica.

A parte II faz uma introdução teórica e alguma prática dos conteúdos, que completa os exemplos ilustrativos, exercícios resolvidos e exercícios propostos com solução que podem ser encontrados nos slides de apoio (Parte III).

Capítulo 1 Análise exploratória de dados

A Estatística (e mais genericamente os métodos quantitativos) tem vindo a revelar-se uma disciplina fundamental, capaz de fornecer métodos e técnicas para recolher, organizar, apresentar, resumir e interpretar os conjuntos de dados, com objetivo de tirar conclusões sobre a informação contida nesses dados e, com base nessas conclusões, poder tomar decisões bem fundamentadas.

As técnicas e os métodos usados para organizar, apresentar, resumir e interpretar um conjunto de dados constitui a metodologia usualmente designada por **estatística descritiva** ou, também, por **análise exploratória de dados.** (Magalhães, Oliveira, & Silva, 2017).

Neste capítulo, desenvolvem-se três temas: o primeiro consiste numa revisão sobre definição de estatística descritiva, população e amostra, variável e tipos de variável, entre outros. O segundo inclui métodos para organização de dados, onde se desenvolve a construção de tabelas e gráficos de frequências. O terceiro tema apresenta medidas amostrais (em particular medidas de localização, dispersão e assimetria, para dados contínuos ou classificados) e formas de visualização de dados (diagrama de barras, histograma e boxplot).

1. Revisão de conceitos de estatística descritiva

A estatística descritiva ocupa-se da organização, condenação e apresentação da informação fornecido por um conjunto de dados, de forma a caracterizar quantitativamente o objeto de estudo. Assim, para além da apresentação dos dados em tabelas de frequência ou recorrendo a métodos gráficos, torna-se necessário caracterizar variáveis por recurso a medidas de resumo, condicionadas pelos níveis de medição.

Perante um conjunto de dados o primeiro passo de uma análise estatística passa por uma análise exploratória de dados que incorpora métodos estatísticos que visam sumariar e descrever os atributos mais proeminentes dos dados, nomeadamente:

- Cálculo numérico de medidas amostrais, tais como valores representativos da tendência central de dados, e valores representativos da quantidade de variabilidade inerente aos dados.
- Resumo e descrição global dos dados através da construção de tabelas e de gráficos.

Segundo a definição apresentada em (Reis, 2009), a estatística descritiva "*consiste na recolha, apresentação, análise e interpretação de dados numéricos através da criação de instrumentos adequados: quadros, gráficos e indicadores numéricos*".

Em estatística, **População** ou **universo** designa um conjunto de unidade individuais com uma ou mais características comuns que se pretendem estudar. A população pode ser finita ou infinita. Exemplos das populações infinitas: os alunos que estudam matemática, os jogadores profissionais de futebol, etc. E, exemplos das populações infinitas: a temperatura em cada ponto da uma cidade, os pontos de uma reta, etc. Um subconjunto finito da população chama-se **amostra**. A

característica dos elementos da amostra que nos interessa averiguar estatisticamente chama-se **variável**.

As variáveis podem ser:

- Variáveis qualitativas ou categorias são as características que definidas por categorias, ou seja, representam uma classificação dos indivíduos. E podem ser nominais ou ordinais.
  - Variáveis nominais: não existe ordenação dentre as categorias. Exemplo: sexo, cor dos olhos, fumador e não fumador, etc.
  - Variáveis ordinais: existe uma ordenação entre as categorias. Exemplos: graus de escolaridade (1º, 2º, 3º graus), estágio da doença (inicial, intermediária, terminal), etc.
- Variáveis quantitativas são características que podem ser descritas por números, sendo estas classificadas entre discretas e contínuas.
  - Variáveis discretas: a variável é avaliada em números que são resultados de contagens e, por isso, somente fazem sentido números inteiros. Exemplos: número de filhos, número de calçados, etc.
  - Variáveis contínuas: a variável é avaliada em números que são resultados de medições e, por isso, podem assumir valores com casas decimais e devem ser medidas por meio de algum instrumento. Exemplos: altura, tempo, idade, etc.

A classificação das variáveis que integram um estudo estatístico é um procedimento necessário e fundamental para a escolha do tipo de procedimentos estatísticos a aplicar na análise dos dados.

2. Organização de dados

A organização de dados permite facilitar a compreensão e a interpretação dos dados e também possibilitar a obtenção de conclusões sobre os dados estudados.

Uma forma simples de organização de dados é através da sua representação numa tabela de distribuição de frequências e/ou num gráfico. Um procedimento que surge naturalmente antes de organizar os dados de forma é o da ordenação dos dados baseada no *rank* (ordem) das observações, quer por ordem acrescente ou decrescente.

2.1. Tabela de frequências

Sejam $x_1^* < x_2^* < \cdots < x_k^*$, de k observações distintas de ordem crescente numa amostra de dimensão n. Geralmente definem-se os seguintes tipos de frequências:

- **Frequência absoluta:** $f_i \equiv$ número de vezes que se observou o valor x_i^* na amostra
- **Frequência relativa:** $f_{ri} = \frac{f_i}{n} \equiv$ proporção de valores iguais a x_i^* na amostra;
- **Frequência absoluta acumulada:** $F_i = f_1 + f_2 + \cdots + f_i = \sum_{j=1}^{i} f_j$
- **Frequência relativa acumulada:** $F_{ri} = f_{r1} + f_{r2} + \cdots + f_{ri} = \sum_{j=1}^{i} f_{rj} = \frac{1}{n}\sum_{j=1}^{i} f_j$

A **tabela de frequência** não é mais do que um quadro que concentra pelo menos um dos tipos de frequências da variável x_i numa amostra ou coleção de dados de dimensão n.

Tabela 1: Distribuições de frequências de variáveis discretas

x_i^*	f_i	F_i	f_{ri}	F_{ri}
x_1^*	f_1	F_1	f_{r1}	F_{r1}
x_2^*	f_2	F_2	f_{r2}	F_{r2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k^*	f_k	n	f_{rk}	1
	$\sum_{i=1}^{k} f_i = n$		$\sum_{i=1}^{k} f_{ri} = 1$	

No caso de distribuição de frequências para variáveis contínuas, obrigam-nos à definição de classes de valores. Para definir estas classes é necessário introduzir alguns novos conceitos: *o número de classes, a amplitude de classes, os limites dos intervalos e o ponto médio ou centro de classe*. Neste caso o objetivo principal de uma distribuição de frequência é encontrar classes com algum significado e utilidade, é de evitar um número muito pequeno ou muito elevado de intervalos. Existem algumas *regras básicas* que deverão ser seguidas na construção dos intervalos (Reis, 2009):

1) Em geral, o número de classes (k) deverá estar compreendido entre quatro e catorze, é adaptada uma das seguintes soluções:
   i. $k = 5$ para $n < 25$
   ii. $k \approx \sqrt{n}$ para $n \geq 25$
   iii. Formula de Sturges: $k = 1 + 3{,}3 \log(n)$
2) Nenhuma classe deverá ter uma frequência nula;
3) As classes deverão ter, sempre que possível, amplitude iguais, e poderá ser calculada por $h = R/k$ em que $R = X_{max} - X_{min}$;
4) Os limites das classes são definidos de modo que cada valor da variável é incluído num e só num intervalo.

Tendo em conta estas regras básicas, muitos autores aceitam que, geralmente, na determinação o número de classe, a **regra de Sturges** fornece bons resultados e que, por isso, é uma boa escolha para iniciar a definição do número de classes a considerar.

Para formalizar a distribuição de frequência para variáveis contínuas ou classificados, sejam os números reais $L_0, L_1, \ldots, L_k$, tais que $L_0 < L_1 < \ldots < L_k$, é definam-se as classes $]L_0, L_1]$, $]L_1, L_2]$, $\ldots$, $]L_{k-1}, L_k]$. Assim, a distribuição de frequências resultante é do tipo:

Tabela 2 : Distribuições de frequências de variáveis contínuas

$\boldsymbol{X}$	$\boldsymbol{f_i}$	$\boldsymbol{F_i}$	$\boldsymbol{f_{ri}}$	$\boldsymbol{F_{ri}}$
$]L_0\,, L_1]$	f_1	F_1	f_{r1}	F_{r1}
$]L_1\,, L_2]$	f_2	F_2	f_{r2}	F_{r2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$]L_{k-1}\,, L_k]$	f_k	n	f_{rk}	1
	$\sum_{i=1}^{k} \boldsymbol{f_i} = n$		$\sum_{i=1}^{k} \boldsymbol{f_{ri}} = 1$	

Nota-se que, o limite superior de uma classe coincide com o limite inferior da classe seguinte. Como as classes têm que ser incompatíveis, é preciso saber onde contar uma observação que tem em valor coincidente com um limite de classe. É usual convencionar que as classes têm o limite superior fechado e o limite inferior aberto, isto é, se classe $]L_{j-1}\,,\ L_j], j = 1, \dots, k$. No entanto, se as classes forem constituídas a partir do valor mínimo da amostra, o limite inferior da classe deve ser fechado e o limite superior deve ser aberto.

Dada uma variável X, a sua **função de distribuição empírica,** também, muitas vezes, designada por **função das frequências acumuladas**, é uma função que indica, para os diferentes valores de x, a frequência (absoluta ou relativa) acumulada até esse valor. (Magalhães et al., 2017). A função de distribuição empírica é definida como sendo uma função real de variável real que, para cada valor real x, indica a frequência (absoluta ou relativa) de observação menores ou iguais a x. Assim, a função de distribuição empírica (frequências relativas acumuladas), é definida por

$$\hat{F}_n(x) = \begin{cases} 0, & se\ x < x_1 \\ F_{r1}, & se\ x_1 \leq x < x_2 \\ F_{r2}, & se\ x_2 \leq x < x_3 \\ \vdots & \\ 1, & se\ x \geq x_k \end{cases}$$

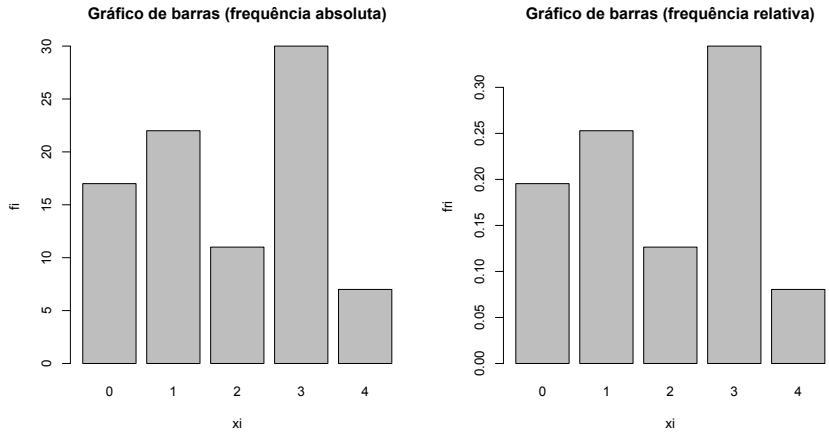
2.2. Representações gráficas da distribuição de frequência

Existem diversos tipos de gráficos, sendo que a escolha do mais apropriado para cada situação depende de vários fatores, como o objetivo da pesquisa ou até mesmo as particularidades das informações a serem apresentadas. Nesta secção iremos apresentar o gráfico de barras, gráfico de linhas, gráfico de setores, histograma e polígono de frequências.

As variáveis qualitativas nominais podem ser bem visualizadas através de gráficos de barras (muitas categorias) e/ou gráfico de setores (poucas categorias). Variáveis qualitativas ordinais, por outro lado, poderão ser melhor visualizadas através de gráficos de barras e gráficos de linhas. Variáveis quantitativas serão melhor ilustradas por intermédio de gráficos de linhas (com ordenação das observações) ou histogramas, polígono de frequências e boxplot (para representação da sua distribuição).

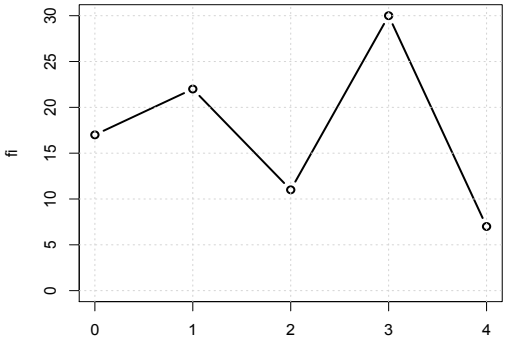
a) **Gráfico de barras** (ou de colunas) é utilizado, em geral, para representar dados de uma tabela de frequências associadas a uma variável qualitativa. Nesse tipo de gráfico, cada barra retangular representa a frequência absoluta ou a frequência relativa da respetiva variável.

Gráfico 1 : Gráfico de barras



b) **Gráfico de linhas** (ou de segmentos) é utilizado, em geral, para representar a evolução dos valores de uma variável no decorrer do tempo.

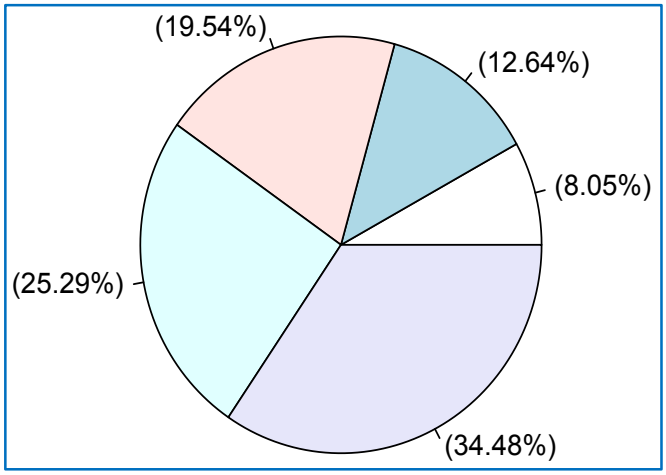
Gráfico 2 : Gráfico de linha



c) **Gráfico de setores**, também conhecido como **"gráfico de pizza"**, ou **"diagrama circular"** é utilizado, em geral, para representar partes de um todo. Para construir um gráfico de setores é necessário determinar o ângulo dos setores circulares correspondentes. Neste caso, consiste em dividir um círculo (360º) em setores proporcionais às realizações de cada categoria através de uma regra de três simples, na qual a frequência total corresponderia aos 360º e a frequência ou a proporção de cada categoria corresponderia a um valor desconhecido em graus.

$$Graus\ de\ uma\ categoria = 360^0 \times \frac{freq.abs.da\ categoria}{dimensão\ da\ amostra} = 360^0 \times \frac{\boldsymbol{f_i}}{\boldsymbol{n}} = 360^0 \times f_{ri}$$

Gráfico 3 : Gráfico de setores

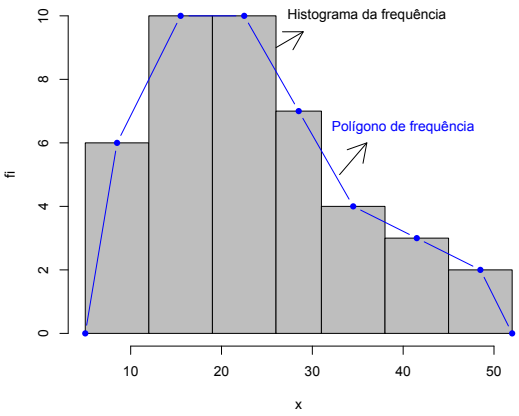


d) Histograma e polígono de frequências.

Histograma é uma representação constituída por uma sucessão de retângulos (barras) adjacentes em que cada um tem por base um intervalo de classe e a altura é igual à respetiva frequência (relativa ou absoluta) dessa classe.

Polígono de frequências é um gráfico de linhas onde são representadas as frequências (relativa ou absoluta) nos pontos médios das classes. Para fechar o polígono basta ligar a frequência associada ao ponto médio da classe extrema ao ponto de abcissa igual ao limite inferior (para a primeira classe) ou ao limite superior (para a última classe) e ordenada zero, tal como ilustrado na figura abaixo.

Gráfico 4 : Histograma e polígono da frequência



3. Medidas amostrais

3.1. Medidas de localização

Medida de localização é um valor que localiza uma dada particularidade do conjunto de dados. No caso particular em que as medidas de localização indicam os valores da variável

estatística onde os dados observados mais se concentram, designam-se por medidas de tendência central. As medidas de tendência central mais usadas são a **média aritmética**, a **moda** e a **mediana**. Existem também outras medidas que medem outras tendências e designam-se por medidas de localização relativa. Exemplos destas medidas são média aparada, mínimo e máximo, e os quantis.

a) **Média aritmética**, muitas vezes designada, apenas por **média** ou **valor médio**. Seja X um conjunto de dados constituído por n observações, $x_1, x_2, \dots, x_n$.

- A média de um conjunto de dados não organizados e não tabelados através de uma distribuição de frequências (Magalhães et al., 2017), é dada por:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Caso os dados observados sejam relativos a uma variável estatística discreta e estejam tabelados, a expressão anterior pode ser, facilmente, adaptada, seja a média dada por (Magalhães et al., 2017):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i {x_i}^* = \sum_{i=1}^{k} f_{ri} {x_i}^*$$

- Para o caso de variáveis contínuas classificadas, o cálculo da média é aproximado uma vez que todas as observações vão ser aproximadas pelo ponto médio da classe à qual pertencem. Assim,

$$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{k} f_i x_{mi} = \sum_{i=1}^{k} f_{ri} x_{mi}$$

onde, $\bar{x}$ é média

n é número de observações

k é número de classes

f_i é frequência absoluta da observação/classe i

f_{ri} é frequência relativa da observação/classe i

$x_{mi} = \frac{L_{inf} + L_{sup}}{2}$ é ponto médio da classe i

L_{inf} e L_{sup} são respetivamente o limite inferior e o limite superior de classe i

b) **Mediana** é o valor da variável que permite dividir o conjunto dos dados ordenados em dois conjuntos de igual tamanho (isto é, 50% dos dados ficam à direita da mediana e 50% dos dados ficam à esquerda da mediana).

A mediana é muito usada, especialmente, quando se considera que os valores extremos têm pouca importância. Com efeito, a mediana é uma medida de localização muito menos sensível do que a média a dados extremos.

Considere-se os dados formado por n observações, $x_1, x_2, \dots, x_n$, represente-se por $x_{j:n}$ o j-ésimo valor do conjunto ordenado dos n dados.

- A mediana de um conjunto de dados de uma variável estatística discreta, cujos estão tabelados, é dada por (Magalhães et al., 2017)

$$M_d = \begin{cases} x_{\frac{n+1}{2}:n} & se\ n\ \text{é}\ impar \\ \dfrac{x_{\frac{n}{2}:n} + x_{\frac{n}{2}+1:n}}{2} & se\ n\ \text{é}\ par \end{cases}$$

- Para o caso de uma variável contínua classificados em classes, o calculo da mediana pode ser assim sumariado (Reis, 2009):
  i) Calcula-se a ordem do valor mediano da amostra ($\frac{n}{2}$ em frequências absolutas ou 50% em frequências relativas). Neste caso, como a variável é contínua não se diferencia entre n par e impar;
  ii) Pelas frequências acumuladas identifica-se a classe que contém a mediana (ordem $\frac{n}{2}$ ou 50% da frequência relativa) e que será a classe mediana;
  iii) Calcula-se o valor aproximado da mediana através da seguinte fórmula:

$$M_d \approx L_{inf} + \frac{\frac{n}{2} - F_{ant}}{f} \times h$$

onde: M_d é mediana

L_{inf} é limite inferior da classe mediana

F_{ant} é frequência absoluta acumulada da classe anterior da classe mediana

f é frequência absoluta da classe mediana

h é amplitude da classe mediana

c) **Média aparada** é a média aritmética após a eliminação de uma certa percentagem de valores extremos, superiores e inferiores, numa amostra de valores ordenados. Assim, aparando 5%, queremos dizer que cortamos 5% dos elementos da cauda esquerda e 5% dos elementos de cauda direita. Nota-se que a média não é mais do que uma média aparada a 0% e a mediana não é mais do que uma média aparada a 50%.

d) **Moda** é o valor da variável estatística mais frequentemente observado, isto é, a moda é o valor da variável que foi observado mais vezes. A moda não tem de ser única pois pode haver mais do que um valor x_i^* com igual frequência sendo essa frequência máxima.

- A moda de um conjunto de dados de uma variável estatística discreta não organizados e não tabelados, é feita por procurando qual o valor que mais vezes repete no conjunto de dados.

- Num conjunto de dados de uma variável estatística discreta organizados através de uma distribuição de frequências, a moda é, formalmente, considere todo o valor da variável estatística que têm maior frequências.
- No caso variável contínua, considere-se a classe de maior frequência como classe modal, e a determinação da moda, é dada por (Reis, 2009)

$$M_o \approx L_{inf} + \frac{\mathrm{d}_1}{\mathrm{d}_1 + \mathrm{d}_2} \times h$$

onde: M_o é moda

L_{inf} é limite inferior de classe modal

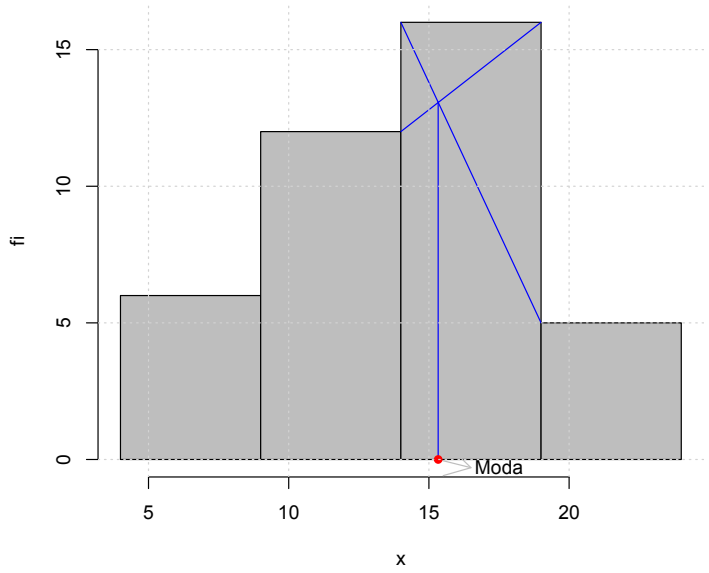
d_1 é diferença da frequência absoluta da classe modal e da classe anterior

d_2 é diferença da frequência absoluta da classe modal e da classe posterior

h é amplitude da classe modal

A moda poderá ser determinada graficamente. Para tal é necessário construir o histograma da distribuição, identificar a classe modal e fazer a seguinte construção:

Gráfico 5 : Ilustração da moda



e) **Mínimo** e **máximo**: as observações mais simples de serem extraídas são o mínimo e o máximo. O mínimo é a observação com *rank* ascendente igual a 1, ou seja, $x_{1:n}$. O máximo é a observação de *rank* ascendente igual a n, corresponde a $x_{n:n}$.

f) **Quantis:** denomina-se por quantil de ordem p, $p \in (0,1)$, o valor real que Q_p que detém, à sua esquerda, (aproximadamente) $p\ x\ 100\%$ das observações que compõem a amostra. Os mais utilizados são os **quartis**, os **decis** e os **percentis**. Os quantis são medidas de localização, mas, na sua maioria, não são medidas de tendência central. O cálculo dos quartis faz-se análogo ao escrito para a mediana.

i) **Quartis,** são os quantis de ordens $p_i = \frac{i}{4}$, com $i = 1,2,3$, ou seja, $p = \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$. Os quartis dividem um conjunto de dados em 4 parte iguais. Assim, o número de quartis é 3, são respetivos 1º quartil ou quartil inferior, 2º quartil ou mediana e 3º quartil ou quartil superior.

ii) **Decis,** são os quantis de ordens $p_i = \frac{i}{10}$, $i = 1, \ldots, 9$, ou seja, $p = \{\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}\}$. Os decis dividem um conjunto de dados em 10 partes iguais.

iii) **Percentis,** são os quantis de ordens $p_i = \frac{i}{100}$, $i = 1, \ldots, 99$, ou seja, $p = \{\frac{1}{100}, \frac{2}{100}, \ldots, \frac{99}{100}\}$. Os percentis, dividem um conjunto de dados em 100 partes iguais.

- Dado um conjunto de dados constituído por n observações, o primeiro passo consiste, sempre, na ordenação dos valores que constituem esse conjunto. Após o conjunto estar ordenado, o quantil de ordem p é dado por (Hall, Neves, & Pereira, 2011)

$$Q_p = \begin{cases} x_{[np]+1:n} & se\ np\ não\ for\ inteiro \\ \frac{1}{2}\left(x_{np:n} + x_{np+1:n}\right) & se\ np\ for\ inteiro \end{cases}$$

onde $[np]$ representa a parte inteira de np.

- No caso variável contínuas e estão tabelados e classificados através de uma distribuição de frequências, classe que contém o quantil pretendido é aquela em que a sua frequência absoluta acumulada contém np, o quantil de ordem p é dada por (Mello, 2014)

$$Q_p \approx L_{inf} + \left(\frac{np - F_{ant}}{f}\right) \times h$$

onde:

Q_p = quantil de ordem p

L_{inf}= Limite inferior de classe quantil de ordem p

n = Total de observações

F_{ant} = Frequência acumulada da classe anterior

h = Amplitude da classe quantil de ordem p

f = Frequência absoluta da classe quantil de ordem p

Do conjunto de definições apresentados conclui-se que a “mediana”, o “segundo quartil”, o “quinto decil” e o “quinquagésimo centil” são exatamente iguais.

3.2. Medidas de dispersão

As medidas de dispersão têm como objetivo descrever a variabilidade ou dispersão existente num determinado conjunto de dados.

a) **Amplitude amostral,** também chamada **amplitude total** ou **amplitude do intervalo de variação,** é diferença entre o máximo e o mínimo. Designa-se usualmente pela letra R devido a palavra inglesa "*Range".*

$$R = x_{max} - x_{min}$$

b) **Amplitude interquartis ou distância interquartis** é a diferença entre o terceiro e o primeiro quartil, isto é,

$$AIQ = Q_{0,75} - Q_{0,25}$$

c) **Variância,** não é mais do que a média dos desvios das observações em relação à média da amostra, depois de elevados ao quadrado por forma evitar cancelamento parcial de diferenças positivas e negativas.

A variância amostral será dada pelo: (Fonseca, 2001);(Magalhães et al., 2017)

- A variância amostral no caso variável discreta

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(x_i^* - \bar{x})^2$$

- A variância amostral no caso variável contínua classificados

$$s^2 \approx \frac{1}{n-1}\sum_{i=1}^{k} f_i(x_{mi} - \bar{x})^2$$

onde, x_i é observação i

n é número de observação

s^2 é variância amostral

$\bar{x}$ é média amostral (a estimar)

x_i^* observação i distinta

k é número de observações distintas (variável discreta) ou número de classe (variável contínuas)

x_{mi} é ponto médio da classe i

d) **Desvio-padrão,** define-se como a raiz quadrada da variância, com a intenção de disponibilizar uma medida de dispersão na mesma unidade dos dados, ao invés do quadrado como sucede com a variância, e naturalmente denotada por s.

3.3. Medidas de assimetria (*skewness*)

Dada a tabela de frequências referente a um conjunto de dados formado por n observações, é, por vezes útil saber se a distribuição das frequências pelos valores possíveis da variável estatística é ou não simétrica e, caso seja assimétrica, será interessante definir uma medida que indica o grau de assimetria e qual o tipo de assimetria existência. Existem várias medidas de assimetria, entre as quais:

a) O coeficiente de assimetria (Pedrosa & Gama, 2016)

$$B = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} \qquad (dados\ não\ classificados)$$

e

$$B = \frac{\frac{1}{n}\sum_{i=1}^{k} f_i(x_{mi} - \bar{x})^3}{s^3} \quad (dados\ classificados\ em\ classe)$$

onde, f_i é frequência absoluta da classe i

x_{mi} é o ponto médio da classe i

A interpretação do coeficiente de assimetria:

- Se $B = 0$, a distribuição é simétrica
- Se $B > 0$, há evidências de assimetria positiva na distribuição
- Se $B < 0$, há evidências de assimetria negativa na distribuição

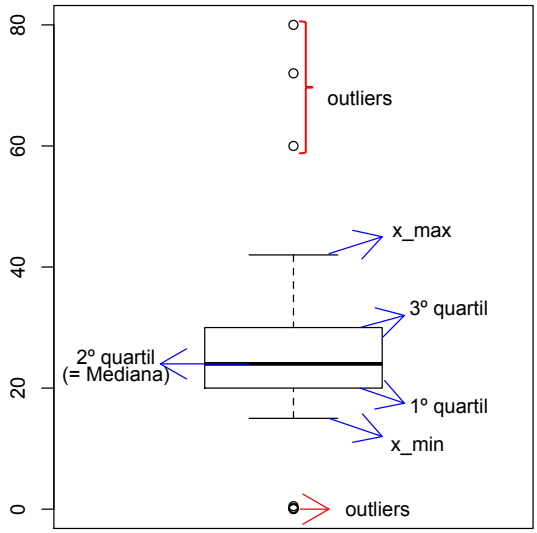
b) Comparando as três medidas de tendência central

- $M_o = M_d = \bar{x} \Longrightarrow$ distribuição simétrica
- $M_o \leq M_d \leq \bar{x} \Longrightarrow$ distribuição assimétrica à esquerda ou assimetria positiva
- $M_o \geq M_d \geq \bar{x} \Longrightarrow$ distribuição assimétrica à direita ou assimetria negativa

3.4. Representação gráfica

Caixas de bigodes (boxplot)**,** é uma maneira de apresentar gráfica e resumidamente algumas medidas de localização e dispersão. O retângulo (caixa) é desenhado de forma a que os seus lados, inferior e superior, correspondam aos 1º e 3º quartis. O segmento no seu interior refere-se à mediana. O mínimo e o máximo do conjunto de dados são representados pelo segmento inferior e superior (bigodes) desenhados no exterior do retângulo. O valor mínimo e máximo representado excluem os valores aberrantes (outliers), que podem ser moderados ou severos.

Gráfico 6 : Boxpot



Estes valores aberrantes correspondem a valores observações que se distanciam bastante do *grosso* das observações. Todas as observações que excedam os limites dos bigodes são identificadas como *outliers.* Habitualmente, são utilizadas as seguintes barreiras para definição de outliers moderados e severos:

- os *outliers moderados* são todas as observações que se situam para além de barreiras $Q_{0,25} - 1{,}5 \times AIQ$ e $Q_{0,75} + 1{,}5 \times AIQ$,
- se, além disso, ainda ultrapassarem as barreiras $Q_{0,25} - 3 \times AIQ$ e $Q_{0,75} + 3 \times AIQ$ os as observações designam-se *outliers severos ou extremos*,

onde $Q_{0,25}$ é primeiro quartil, $Q_{0,75}$ é terceiro quartil e AIQ é amplitude interquartil (Hall et al., 2011).

Assim, o *boxplot* espelha de forma simples a estrutura da população subjacente aos dados e nele é possível identificar as seguintes características evidenciadas pelos dados:

- Localização: indicada pela mediana,
- Dispersão: resultante do comprimento da caixa (distância interquartis AIQ) e do comprimento total entre os extremos dos bigodes (amplitude R),
- Comprimento das caudas da distribuição subjacente: dado pelo comprimento das linhas que definem os bigodes,
- Assimetria da distribuição: indicada pela assimetria da caixa de bigodes.

Capítulo 2 Análise de tabelas de contingência

Neste capitulo, irão desenvolver-se cinco temas. O primeiro diz respeito ao conceito geral de tabela de contingência. O segundo consiste numa revisão sobre espaços amostrais e eventos, probabilidade condicionada, eventos independentes, disjunção e entre outros que vão utilizar para compreender melhor os conteúdos seguintes. O terceiro inclui teste de hipóteses, desenvolvendo-se sobre teste de qui-quadrado (χ^2) de independência, homogeneidade e ajustamento; teste de razão verossimilhanças e os testes alternativa para as tabelas de contingência 2×2, isto é teste exato de Fisher para amostras independentes e teste McNemar para amostras emparelahadas ou correlacionadas. O quarto tema aborda tabelas de contingência bidimensional, onde se desenvolvem técnicas para análise de resíduos com objetivo para encontrar a localização de fontes de dependência; modelos log-lineares de independência e de saturado ou modelo completo. O último tema incide sobre tabelas de contingência tridimensional, e neste tema desenvolvem-se os modelos de log-lineares, considerando o modelo de independência: mútua, parcial,condicional ou associação 2 a 2 e o modelo saturado.

1. Tabelas de contingência

Suponha que de uma amostra aleatória de tamanho *n,* de uma dada população, são observadas duas características A e B (qualitativas ou quantitativas), com r e c categorias, respetivamente $A_1, \dots, A_r$ e $B_1, \dots, B_c$.

Cada individuo da amostra é classificado pelo cruzamento de duas categorias. Podemos representar as frequências observadas (O_{ij}, $\text{com } i = 1, \dots, r;\ j = 1, \dots, c$) numa tabela de dupla entrada, a que se chama *tabela de contingência* $\boldsymbol{r \times c}$, como a que se observa de seguida:

Tabela 3 : Tabela de contingência

		$\boldsymbol{B}$			**Totais**
		$\boldsymbol{B_1}$	...	B_c	
$\boldsymbol{A}$	$\boldsymbol{A_1}$	O_{11}	...	O_{1c}	$\sum_{j=1}^{c} O_{1j} = \boldsymbol{O_{1.}}$
	⋮	⋮		⋮	⋮
	$\boldsymbol{A_r}$	O_{r1}	...	O_{rc}	$\sum_{j=1}^{c} O_{rj} = \boldsymbol{O_{r.}}$
Totais		$\sum_{i=1}^{r} O_{i1} = \boldsymbol{O_{.1}}$	...	$\sum_{i=1}^{r} O_{ic} = \boldsymbol{O_{.c}}$	$\sum_{i=1}^{r}\sum_{j=1}^{c} O_{rc} = \boldsymbol{n}$

Assim, a tabela de contingência é uma representação de dados, quer do tipo qualitativo ou quer do tipo quantitativo que podem ser classificados segundo dois critérios. Nesta tabela, as linhas correspondem a um dos critérios e as colunas correspondem ao outro critério. No interior da tabela, as células correspondem ao número de observações O_{ij}, que satisfazem ambos os critérios.

2. Revisão de conceitos de probabilidade

Considerem-se os seguintes elementos:

a) Espaço amostral ou espaço dos resultados, é o conjunto de todos os resultados possíveis de ocorrência de um evento, simbolicamente representada pela: $\boldsymbol{S}$ ou $\boldsymbol{\Omega}$. O número de elementos do espaço amostral é denotado por $\boldsymbol{n}(\boldsymbol{S})$ ou $\boldsymbol{n}(\boldsymbol{\Omega})$ e $\boldsymbol{n}(\boldsymbol{\Omega}) \neq \boldsymbol{0}$.

b) Evento é qualquer subconjunto de um espaço amostral, denotada por qualquer uma letra maiúscula. O número de elementos de um evento é denotado por $\boldsymbol{n}(.)$.

c) Probabilidade de um acontecimento (evento) é quociente entre o número de casos favoráveis ao acontecimento e o número de casos possíveis, supondo que todos os casos são igualmente possíveis. Ou seja:

$$\boldsymbol{P}(.) = \frac{\sharp\ casos\ favoráveis}{\sharp\ casos\ possíveis} = \frac{\boldsymbol{n}(.)}{\boldsymbol{n}(\boldsymbol{\Omega})}, \quad 0 \leq P(.) \leq 1$$

d) Classificação de eventos

- Evento certo, quando ele possui todos os elementos do espaço amostral. Ou seja, $\boldsymbol{n}(.) = \boldsymbol{n}(\boldsymbol{\Omega})$. Neste caso, $\boldsymbol{P}(.) = \boldsymbol{1}$
- Evento impossível, quando número de casos favoráveis é zero. ou seja, $\boldsymbol{n}(.) = \boldsymbol{0} \Longrightarrow P(\emptyset) = 0$
- Evento complementar, dado um evento A num espaço amostral $\boldsymbol{\Omega}$. O complementar do evento $\bar{A}$ são todos os elementos do espaço amostral $\boldsymbol{\Omega}$ que não estão contidos em A, então temos que $\bar{A} = \boldsymbol{\Omega} - \boldsymbol{A}$ e ainda $\boldsymbol{\Omega} = \bar{A} + \boldsymbol{A}$.

  A probabilidade de evento complementar é:

  $$P(\bar{A}) = P(\boldsymbol{\Omega} \backslash A) = P(\boldsymbol{\Omega}) - P(A) = 1 - P(A)$$

- Evento união. Dados dois eventos A e B de um espaço amostral $\boldsymbol{\Omega}$. O número de elementos de $A \cup B$ é igual à soma do número de elementos de A com o número elemento de elemento de B, menos uma vez o número de elementos de $A \cap B$ que foi contado duas vezes, assim temos:

  $$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

  A probabilidade de ocorrência A ou B é dada por:

  $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

  Se A e B são dois eventos disjuntos ou mutuamente exclusiva. Ou seja,

  $$A \cap B = \emptyset \Longrightarrow P(A \cap B) = 0$$

  Logo, a sua probabilidade é simplificada por:

$$P(A \cup B) = P(A) + P(B)$$

- Evento intersecção. Dados dois eventos A e B de um espaço amostral $\boldsymbol{\Omega}$. O número de elementos de $A \cap B$ é igual o número de elementos simultâneos em A e B.

  Assim, a probabilidade da intersecção de dois eventos $P(A \cap B)$ ou probabilidade conjunta A e B corresponde à possibilidade dos dois eventos ocorrem simultânea ou sucessivamente.

  Para calculo de $P(A \cap B)$ pode ser feito à custa das probabilidades condicionadas $P(A|B)$ ou $P(B|A)$, onde $P(A|B)$ representa a probabilidade de ocorrência de A sabendo que B ocorreu. A relação entre as probabilidades é a seguinte

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \Leftrightarrow \quad P(A \cap B) = P(B) \times P(A|B)$$

  ou

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \Leftrightarrow \quad P(A \cap B) = P(A) \times P(B|A)$$

  Se A e B são eventos independentes, a probabilidade de ocorrência simultânea de A e B, é simplificada por:

$$\left.\begin{matrix} P(A|B) = P(A) \\ P(B|A) = P(B) \end{matrix}\right\} \Leftrightarrow \quad P(A \cap B) = P(A) \times P(B)$$

  pois, se A e B são independentes, $P(A|B)$ é igual a $P(A)$ já que a probabilidade de ocorrência de A não depende se B ocorreu.

3. Testes de hipóteses

3.1. Teste de independência de Qui-quadrado e de razão verosimilhanças

Quando estudamos duas características diferentes da mesma população, é frequente querermos verificar se são independentes, isto é, se não existe nenhuma relação entre elas. Por exemplo, será que a opinião acerca do casamento é independente da idade? Será que a situação de emprego depende do sexo? Será o peso é independente da altura de uma pessoa? Em qualquer destas situações estão presentes duas variáveis (por exemplo, situação de emprego e sexo) que são classificadas. Nestes casos, precisamos de realizar teste de independência estatística.

a) Teste Qui-quadrado de independência

O teste de Qui-quadrado, proposto por Karl Pearson em 1900, é um teste não paramétrico que se aplica a amostra independentes. É também designado por “teste qui-quadrado de Pearson” e simbolicamente denotado por χ^2, servindo para testar hipóteses estatísticas de independência. O

objetivo deste teste de independência é testar se duas variáveis, expressas numa tabela de contingência, são independentes entre si.

Pressuposto de aplicação do teste Qui-quadrado (Mello, 2014):

- $n \geq 20$
- Todos os frequências esperadas (E_{ij}) forem superiores a 1
- Pelo menos 80% dos E_{ij} forem não inferiores a 5

O teste do Qui-quadrado é aplicável se as condições acima mencionadas são satisfeitas. Caso contrário, a distribuição associada à estatística de teste é desconhecida, pelo que não se sabe como decidir sobre o H_0.

Hipóteses a testar:

H_0: as variáveis são independentes (não estão associadas).

H_1: as variáveis não são independentes.

Assim, pretendem-se comparar as frequências observadas (O_{ij}) de cada uma das $r \times c$ células, com as correspondentes frequências esperadas (E_{ij}) supondo H_0 verdadeiro.

Como obter frequências esperadas ($\mathbf{E_{ij}}$)

Se H_0 for verdadeiro, isto é, as variáveis A e B forem independentes então:

$$E_{ij} = n \times P(A_i \cap B_j) = n \times P(A_i) \times P(B_j) = n \times \frac{O_{i.}}{n} \times \frac{O_{.j}}{n} = \frac{O_{i.} \times O_{.j}}{n}$$

Pois a probabilidade de uma intersecção é igual ao produto das probabilidades.

Portanto, as hipóteses deste teste são:

$H_0: O_{ij} = E_{ij}, ; \forall_{ij}$

$H_1: O_{ij} \neq E_{ij}$; para algum i, j

Estatística do teste (Mello, 2014)

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

onde, O_{ij} = frequências observadas

E_{ij} = frequências esperadas

Quando o número de observações é elevado, a estatística χ^2 é aproximadamente a Qui-quadrado com $(r-1) \times (c-1)$ graus de liberdade (isto é, $\chi^2 \sim \boldsymbol{\chi^2_{(r-1)\times(c-1)}}$).

Rejeita ou não se rejeita a hipótese nula de que a distribuição de frequências observadas é a mesma que a distribuição da frequências esperadas baseada em comparação de estatística de teste e o valor crítico de χ^2. Neste caso, rejeita hipótese nula se o valor do teste estatístico excede ou igual o valor crítico de χ^2 num determinado nível de significância α. Caso contrário, não se

rejeitar hipótese nula. Ou seja, rejeita-se H_0 (em favor de H_1) quando χ^2_{obs} é maior ou igual ao valor crítico do teste, seja $\boldsymbol{\chi^2_{(1-\alpha);(r-1)\times(c-1)}}$. Se $\chi^2_{obs} < \boldsymbol{\chi^2_{(1-\alpha);(r-1)\times(c-1)}}$ então não se rejeita H_0.

b) Medidas de associação

No teste do Qui-quadrado apresentado, se H_0 de independência for rejeitada, pode interessar medir a intensidade da associação entre duas variáveis, através de uma medida de associação adequada.

O valor de uma medida de associação é geralmente baseada no valor de estatística do teste de qui-quadrado, χ^2_{obs}. Assim, se $\chi^2_{obs} = 0$ então a medida de associação deve ser igual ao zero (situação de independência). Além disso, quanto maior for a medida de associação, maior o grau de dependência entre as duas variáveis.

(1) **Coeficiente de contingência de Pearson (Mello, 2014)**

$$C = \sqrt{\frac{\chi^2_{obs}}{\chi^2_{obs} + n}}, 0 \leq C \leq \sqrt{\frac{min\{r, c\} - 1}{min\{r, c\}}}$$

onde, C é coeficiente de contingência de Pearson

χ^2_{obs} é estatística do teste qui-quadrado

n é dimensão da amostra

r é número de linha de tabela contingência

c é número de coluna de tabela de contingência

(2) **Coeficiente de Tscprow (Mello, 2014)**

$$T = \sqrt{\frac{\chi^2_{obs}}{n\sqrt{(r-1)(c-1)}}}, 0 \leq T \leq 1, \text{onde } T = 1 \text{ ocorre apenas para } r = c.$$

(3) **Coeficiente de contingência de V de Cramer (Mello, 2014)**

$$V = \sqrt{\frac{\chi^2_{obs}}{n[min\{r, c\} - 1]}}, 0 \leq V \leq 1$$

(4) **Coeficiente fi (Mello, 2014)**

É uma medida de associação de duas variáveis formado uma tabela de contingência 2×2.

Tabela 4 : Tabela de contingência 2 × 2

		Variável x_1		**Total**
		1	**2**	
Variável x_2	**1**	a	b	$\boldsymbol{a+b}$
	2	c	d	$\boldsymbol{c+d}$
Total		$\boldsymbol{a+c}$	$\boldsymbol{b+d}$	$\boldsymbol{n}$

Supondo H_0 de independência verdadeira, a estatística de teste de χ^2 é

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

e, normalizando de forma adequada, obtem-se, $\varphi = \sqrt{\frac{\chi^2_{obs}}{n}} = \frac{|ad-bc|}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$, $0 \leq \varphi \leq 1$ que pode ser usuada como uma medida de associação. Quanto maior for o valor de φ maior a associação entre as variáveis.

Alternativamente, pode-se utilizar $\varphi' = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$, $-1 \leq \varphi' \leq 1$ que permite medir a intensidade da associação e também a sua direção

- $\varphi' = 0 \rightarrow ad - bc = 0 \Longrightarrow$ ausência de associação (i.e., independência)
- $\varphi' > 0 \rightarrow ad > bc \Longrightarrow$ associação positiva
- $\varphi' < 0 \rightarrow ad < bc \Longrightarrow$ associação negativa

c) Teste de razão verosimilhanças

O teste de razão de verosimilhanças é semelhante do teste χ^2 de Pearson e é denotado por G^2 (Howell, 2000). O teste de G^2 baseia-se na teoria da máxima verosimilhança. E, cada um desses testes tem algumas vantagens e algumas desvantagens. E, Os resultados destes dois testes geralmente são muito semelhantes, isto é à medida que o tamanho das amostras aumenta, as duas estatísticas do teste convergem.

O uso do teste G de ajustamento é para uma variável nominal com dois ou mais valores (por exemplo, masculinas e femininas, ou vermelhas, rosa e brancas). comparar as frequências observadas em cada categoria com as frequências esperadas. Se o número esperado de observações em qualquer categoria for muito pequeno, o teste G pode dar resultados inexato e, em vez disso, deve ser usar um teste exato. (McDonald, n.d.)

O teste χ^2 e G^2 foram commumente usadas como o teste de ajustamento e teste de independência numa tabelas de contingência e análise multivariada. (Eyduran & Unit, 2005)

Uma vantagem do cálculo de verosimilhança é que o G^2 para uma tabela de dimensões elevadas pode ser cuidadosamente decomposto em componentes menores. Isso não pode ser feito exatamente com o qui-quadrado de Pearson, e o G^2 é a estatística usual para análises log-lineares.

Hipóteses a testar

H_0: as variáveis são independentes

H_1: $\sim H_0$

Se H_0 for verdadeira, as frequências esperadas são estimadas por:

$$E_{ij} = \frac{o_{i.} \times o_{.j}}{n}$$

Estatística de teste

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c}\left(O_{ij}\ln\left(\frac{O_{ij}}{E_{ij}}\right)\right) \sim \chi^2_{(r-1)\times(c-1)}$$

Quando o número de observações é elevado, a estatística G^2 é aproximadamente a Qui-quadrado com $(r-1)\times(c-1)$ graus de liberdade (isto é, $\chi^2 \sim \boldsymbol{\chi^2_{(r-1)\times(c-1)}}$). Assim, tal como no teste de χ^2, rejeita-se H_0 em favor de H_1 quando $G^2 \geq \chi^2_{(1-\alpha);(r-1)\times(c-1)}$. Caso contrário não se rejeita H_0 em favor de H_1.

3.2. Outro teste de Qui-quadrado

Além de teste de Qui-quadrado de independência, existe outros testes de Qui-quadrado que podem ser aplicados a tabelas $\boldsymbol{r \times c}$, são "teste de homogeneidade" e "teste de ajustamento ou teste de aderência" que embora testem hipóteses diferentes têm cálculo semelhantes de teste de Qui-quadrado de independências.

a) Teste qui-quadrado de homogeneidade

Objetivo: comparar a distribuição de contagens para dois ou mais grupos usando a mesmas variáveis categóricas. Por exemplo testar se as populações em A (variável linha numa tabela de contingência) são homogéneas.

Hipóteses a testar

H_0: as proporções de $A_1, \ldots, A_r$ são iguais para todas as categorias de B (isto é, as populações são homogéneas).

$H_1: \sim H_0$

Os dois testes distinguem-se pela forma de como as amostras são recolhidas. Tipicamente, num teste de homogeneidade, fixam-se os totais marginais para populações.

Como funciona o teste de homogeneidade

A realização deste teste é semelhante à do teste de independência, isto é, as hipóteses deste teste são equivalentes às do teste de independência. Isto é,

$H_0: O_{ij} = E_{ij}; \forall_{ij}$

$H_1: O_{ij} \neq E_{ij}$; para algum i, j

Estatística do teste

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \sim \boldsymbol{\chi^2_{(r-1)\times(c-1)}}$$

Decisão: rejeita-se H_0 (em favor de H_1) quando $\chi^2_{obs} \geq \boldsymbol{\chi^2_{(1-\alpha);(r-1)\times(c-1)}}$. Senão então não se rejeita H_0.

b) Teste qui-quadrado de ajustamento

Objetivo: testar se as observações seguem uma determinada distribuição teórica de probabilidade (discreta ou contínua, com ou sem parâmetros conhecidos).

Hipóteses a testar

H_0: A população segue uma determinada distribuição $\mathcal{D}$

H_1: A população não tem uma determinada distribuição $\mathcal{D}$

Como funciona o teste de ajustamento

Para a realização do teste, os dados têm que estar agrupados em k classes (intervalos ou categorias). No caso em que a distribuição $\mathcal{D}$ é contínua, tais classes podem ser baseadas nas classes do histograma.

Neste teste também são comparadas duas quantidades:

- O número de observados em cada categoria (frequência observada, O_i)
- O número de valores que se teriam em cada categoria admitindo que a população tem a distribuição $\mathcal{D}$ (frequência esperada, E_i).

Tabela 5 : Tabela de frequências observadas e frequências esperadas

	$\boldsymbol{C_1}$ ... $\boldsymbol{C_k}$	**Total**
Frequências observadas	O_1 ... O_k	$\boldsymbol{n}$
Frequências esperadas	E_1 ... E_k	$\boldsymbol{n}$

Assim, as hipóteses deste teste são:

$H_0: O_i = E_i \,; \forall_i$

$H_1: O_i \neq E_i$; para algum i

A frequência esperada de uma classe, quando H_0 é verdadeira, é dada por:

$$E_i = n \times P_i \text{ com } P_i = P(C_i)$$

Estatística de teste

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \boldsymbol{\chi^2_{(1-\alpha);(k-m-1)}}$$

Quando o número de observações é elevado, a estatística χ^2 tem aproximadamente distribição do Qui-quadrado com $(k - m - 1)$ graus de liberdade, onde:

- k representa o número de categorias e
- m representa o número de parâmetros de $\mathcal{D}$ que é necessário estimar a partir da amostra.

Decisão: rejeita-se H_0 (em favor de H_1) quando $\chi^2_{obs} \geq \boldsymbol{\chi^2_{(1-\alpha);(k-m-1)}}$. Senão então não se rejeita H_0.

3.3. Teste alternativo de independência para tabelas 2×2

a) Teste exato de Fisher

É um teste não paramétrico adequado para amostras independentes de pequena dimensão e relativamente a uma variável dicotómica. Este teste é uma alternativa ao teste qui-quadrado no caso de tabelas 2×2 (ver tabela 5), quando os pressupostos são violados para sua aplicação.
Este teste permite a calcular a probabilidade exata de ocorrência de uma frequência observada, ou de valor mais extremos (p_{-Value}).
Tal como no teste de qui-quadrado de independência, as hipóteses a testar são:

H_0: as variáveis são independentes

H_1: as variáveis não são independentes

quando H_0 é verdadeira, então $E_{ij} = \frac{O_{i.} \times O_{.j}}{n}$; $\quad i = \{1,\ 2\},\ j = \{1,\ 2\}$.

Como funciona o teste (Mello, 2014)

Se $O_{11} > E_{11}$ então a associação é positiva (cauda da direita)

- Hipóteses a testar são

  H_0: independência

  H_1: associação positiva
- Construir as tabelas de frequências de observadas (semelhante com tabela 4), mantendo os mesmos totais marginais e aumentando a frequência observada $O_{11} = a$ até ao menor total marginal que lhe corresponde.
- Calcular a probabilidade de observada p_a com $a = O_{11}$ para cada tabela construída.

$$p_a = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$$

- Calcular a probabilidade de significância p, é dado pela soma das probabilidades calculadas para cada tabela.

$$p_{-Value} = p_a + p_{a+1} + \cdots + p_{min\,(a+b,a+c)} = p_{sup}$$

Se $O_{11} < E_{11}$ então a associação é negativa (cauda da esquerda)

- Hipóteses a testar são

  H_0: independência

  H_1: associação negativa
- Construir as tabelas de frequências de observadas mantendo os mesmos totais marginais e decrescendo a frequência observada $O_{11} = a$ até zero.
- Calcular a probabilidade de observada p_a para cada tabela construída.
- Calcular a probabilidade de significância p, é dado pela soma das probabilidades calculadas para cada tabela.

$$p_{-Value} = p_a + p_{a-1} + \cdots + p_0 = p_{inf}$$

Regra de decisão

- Teste unilateral: rejeita H_0 em favor de H_1 se $P_{\_Value} \leq \alpha$. Caso contrario não se rejeita H_0 em favor de H_1.
- Teste bilateral: rejeita H_0 em favor de H_1 se $p_{sup} \leq \frac{\alpha}{2}$ ou $p_{inf} \leq \frac{\alpha}{2}$. Caso contrario não se rejeita H_0 em favor de H_1.

b) Teste McNemar

O teste McNemar é aplicado para as amostras emparelhadas e quando os dados aparecem em escala nominal categorizada e são obtidos em dois tempos ou situações distintas e consecutivas (do tipo "**antes e depois**"). Os dados são organizados em tabela de contingência 2×2 em que as frequências observadas em cada célula representam o número de sujeitos que mudaram ou não mudaram de condições.

Tabela 6 : Tabela de contingência 2 × 2 - antes e depois

		Depois		Totais
		+	-	
Antes	+	a	b	$\boldsymbol{a+b}$
	-	c	d	$\boldsymbol{c+d}$
Totais		$\boldsymbol{a+c}$	$\boldsymbol{b+d}$	$\boldsymbol{n}$

Da tabela observa-se que

- a e d é o número de indivíduos que não mudaram de condição,
- b é o número de insucessos, são indivíduos que mudaram de (+) para (-),
- c é o número de sucessos, são indivíduos que mudaram de (-) para (+),
- $b + c$ é o total de indivíduos que mudaram de condição.

Hipóteses a testar

H_0: não existe diferença antes e depois do tratamento

H_1: existe diferença antes e depois do tratamento

Considere-se $n = b + c$ e α o nível de significância. A escolha do teste e a regra de decisão associada podem ser resumidas do modo que se segue: (Reis, Melo, Andrade, & Calapez, 2016)

- Se $b + c \leq 20$, aplica-se o teste binomial

  $P_{\_Value} = P[X = x] = \binom{n}{x} P^x (1-P)^{n-x}$ onde $X \sim B(b + c, \frac{1}{2})$, com $x = min\ \{b, c\}$.

- Se $b + c > 20$, usa-se o teste de χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(b-c)^2}{b+c} \sim \boldsymbol{\chi^2_{(1)}}$$

Demonstração:

$$E_i = np_i = (b+c)\times\frac{1}{2} \Leftrightarrow E_i = \frac{b+c}{2}$$

$$\chi^2 = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i} = \frac{\left(b-\frac{b+c}{2}\right)^2}{\frac{b+c}{2}} + \frac{\left(c-\frac{b+c}{2}\right)^2}{\frac{b+c}{2}}$$

$$= \frac{b^2 - 2b\left(\frac{b+c}{2}\right) + \left(\frac{b+c}{2}\right)^2 + c^2 - 2c\left(\frac{b+c}{2}\right) + \left(\frac{b+c}{2}\right)^2}{\frac{b+c}{2}}$$

$$= \frac{b^2 - b^2 - bc + \left(\frac{b+c}{2}\right)^2 + c^2 - bc - c^2 + \left(\frac{b+c}{2}\right)^2}{\frac{b+c}{2}} = \frac{-2bc + 2\frac{(b+c)^2}{4}}{\frac{b+c}{2}}$$

$$= \frac{-2bc + \frac{(b+c)^2}{2}}{\frac{b+c}{2}} = \frac{\frac{-4bc + (b+c)^2}{2}}{\frac{b+c}{2}} = \frac{-4bc + (b+c)^2}{2}\times\frac{2}{b+c}$$

$$= \frac{-4bc + b^2 + 2bc + c^2}{b+c} = \frac{b^2 - 2bc + c^2}{b+c} = \frac{(b-c)^2}{b+c}$$

Decisão: rejeitar H_0 em favor de H_1 se $P_{\_Value} \leq \alpha$. Caso contrário não se rejeitar H_0 em favor de H_1.

4. Tabelas de contingência $r \times c$

Uma tabela contingência, com r linhas e c colunas diz-se que tem dimensão $r \times c$, e designa-se por tabela *bidimensional* (ver tabela 3).

4.1. Localização de fontes de dependência por análise de resíduos

No caso de rejeitar H_0 de independência, a localização de fontes de dependência indica-nos quais as células que mais contribuem para a dependência. Esta localização pode ser feita através de uma "**análise dos resíduos".**

O processo utilizado para identificação das categorias responsáveis por um valor significante da estatística qui-quadrado foi sugerido por (Haberman, 1973). Este processo envolve a análise dos resíduos padronizados: ("Análise de Resíduos - Tabela Cruzada | Portal Action," n.d.)

$$R_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}\left(1-\frac{O_{i.}}{n}\right)\left(1-\frac{O_{.j}}{n}\right)}}$$

Se H_0 de independência for verdadeira, então $R_{ij} \sim N\ (0,1)$ quando $n \to \infty$.

Comparando $|R_{ij}|$ com o quantil de probabilidade $Z_{1-\frac{\alpha}{2}}$ da distribuição normal reduzida:

- As células tais que$|R_{ij}| \geq Z_{1-\frac{\alpha}{2}}$, contribuem (de forma significativa) para a dependência das variáveis,
- e quanto maior for o valor de $|R_{ij}|$ maior é a contribuição dessa parcela para a dependência.

4.2. Modelos log-lineares

A análise log-linear de tabelas de contingência permite averiguar a existência, ou não, de independência entre as variáveis e quantificar os efeitos que as várias variáveis ou combinações destas exerceram sobre os valores observados.

O objetivo principal desta análise consiste, essencialmente, em ajustar aos dados modelos que reflitam, tão bem quanto possível, a estrutura que lhe foi imposta.

A distribuição de Poisson é uma das mais populares na aplicação de modelos log-lineares para tabelas de contingência, uma vez que as frequências numa tabela são contagens de efectivos de classes, que assumem valores inteiros não negativos.

Vamos considerar o modelo Poisson para tabelas de contingência bidimensional. No entanto, os conceitos introduzidos aplicam-se igualmente aos casos mais gerais. Neste modelo, o tamanho total da amostra não é fixo e todas as contagens são, portanto, aleatórias. Então, as frequências observadas O_{ij} $(i = 1, \dots, r; j = 1, \dots c)$ não são mais do que realizações independentes das variáveis aleatórias O_{ij} bem modelados por uma distribuição de Poisson com valor esperado $E(O_{ij}) = var\,(O_{ij}) = E_{ij}$ $(E_{ij} > 0 \text{ com } i = 1, \dots, r; j = 1, \dots c)$. Assim,

$$\{O_{ij}\} \sim Poisson\,(\{E_{ij}\}).$$

Admitindo O_{ij} $(i = 1, \dots, r; j = 1, \dots c)$ independentes e identicamente distribuídas (i.i.d.), a função massa de probabilidade conjunta para O_{ij} condicionada a E_{ij} é dada por

$$f(\{O_{ij}\}|\{E_{ij}\}) = \prod_{i=1}^{r}\prod_{j=1}^{c} f(O_{ij}|E_{ij}) = \prod_{i=1}^{r}\prod_{j=1}^{c} \frac{\exp(-E_{ij}) E_{ij}^{O_{ij}}}{O_{ij}!}$$

com $O_{ij} \in \mathbb{N}_0$ *e* $E_{ij} \in \mathbb{R}^+$. Esta constitui a função de verosimilhança usual de Poisson e, maximizando o logaritmo desta função obtêm-se os estimadores de máxima verosimilhança para E_{ij}. Maior detalhe sobre o procedimento de obtenção de estimadores pode ser consultado em ("8 - Modelos Log Lineares para Tabelas de Contingência - Tabela Cruzada | Portal Action," n.d.).

A expressão de um modelo refere-se a uma estrutura conceptual ajustável às observações. Portanto, o modelo adequado aos dados é aquele que, apresentando um bom ajustamento aos dados, seja interpretável e possua número mínimo de parâmetros. Estes modelos relacionam as probabilidades associadas a cada célula com uma função de vários parâmetros, a logaritmação das probabilidades permite essa linearização.

Neste caso, vamos considerar dois modelos log-lineares: o modelo independência e o modelo saturado. Um modelo com o número de parâmetros igual ao número de células da tabela é

designado por *saturado*. Estes modelos possuem um ajustamento perfeito já que, nesta situação, as frequências esperadas são, simplesmente, as próprias frequências observadas. A eliminação de termos do modelo saturado conduz a modelos *reduzidos* ou não saturados.

O ajustamento de um dado modelo a uma tabela far-se-á com base em testes estatísticos de determinação hipóteses de independência, isto, é, utilizando as estatísticas qui-quadrado, χ^2 ou razão verossimilhança, G^2 (Leal, 1997). Uma vez ajustado o modelo, os estimadores dos seus parâmetros permitir-nos-ão quantificar os efeitos que as diversas variáveis e as interações entre elas exerceram sobre os dados e analisar a sua significância.

Neste trabalho, aborda-se a problemática da adaptação dos modelos log-lineares à ordinalidade das variáveis em tabela de contingência. Neste sentido, são abordados os modelos log-lineares ordinais que permitem descrever padrões de associação e interação (ou a sua ausência) inerentes à ordinalidade das variáveis. Nestes métodos ordinais usam a informação relativa à hierarquia das categorias, sendo exposta a vantagem que dai advém nomeadamente no que respeita à simplificação de modelos.

a) Modelo independência

No modelo de independência, a probabilidade conjunta é o produto das marginais. Como apresentado anteriormente, duas variáveis categorizadas, organizadas numa tabela bidimensional, dizem-se estatisticamente independentes se,

$$P(A_i \cap B_j) = P(A_i) \times P(B_j) = \frac{O_{i.}}{n} \times \frac{O_{.j}}{n}, i = 1,..,r \text{ e } j = 1,\dots,c$$

Supondo H_0 de independência verdadeira, a frequência esperada é dada por

$$E_{ij} = nP(A_i \cap B_j) = n\frac{O_{i.}}{n} \times \frac{O_{.j}}{n} = \frac{O_{i.} \times O_{.j}}{n}$$

Assim, o modelo de logaritmo é dado por:

$$\ln E_{ij} = \ln\left(\frac{O_{i.} \times O_{j.}}{n}\right) = \ln O_{i.} + \ln O_{.j} - \ln n$$

Isto é, se as variáveis forem independentes, o logaritmo natural da frequência esperada E_{ij} é a soma de

- efeito linha i (variável A),
- efeito coluna j (variável B),
- efeito constante.

Assim, o modelo log-linear de independência deve ser escrito da seguinte forma:

$$\ln E_{ij} = \mu + \lambda_i^A + \lambda_j^B \quad \text{..... (*)}$$

Para estimar os parâmetros do modelo é necessário considerar

$$\sum_{i=1}^{r} \lambda_i^A = 0 \qquad e \qquad \sum_{j=1}^{c} \lambda_j^B = 0,$$

e assim os parâmetros do modelo são estimados através de

- $\hat{\mu} = \frac{1}{r \times c} \sum_{i=1}^{r} \sum_{j=1}^{c} \ln E_{ij}$; efeito médio global

- $\hat{\lambda}_i^A = \frac{1}{c}\sum_{j=1}^{c} \ln E_{ij} - \mu$; efeito principal da variável A
- $\hat{\lambda}_j^B = \frac{1}{r}\sum_{i=1}^{r} \ln E_{ij} - \mu$; efeito principal da variável B

Demonstração:

$$\ln E_{ij} = \ln O_{i.} + \ln O_{.j} - \ln n \Leftrightarrow \ln E_{ij} = \ln E_{i.} + \ln E_{.j} - \ln n \quad \text{..... (i)}$$

somando respetivamente em i, j e i e j, obtivemos

- $\sum_{i=1}^{r} \ln E_{ij} = \sum_{i=1}^{r} \ln E_{i.} + \sum_{i=1}^{r} \ln E_{.j} - \sum_{i=1}^{r} \ln n$

$$\Leftrightarrow \sum_{i=1}^{r} \ln E_{ij} = \sum_{i=1}^{r} \ln E_{i.} + r \ln E_{.j} - r \ln n$$
$$\Leftrightarrow \frac{1}{r}\sum_{i=1}^{r} \ln E_{ij} = \frac{1}{r}\sum_{i=1}^{r} \ln E_{i.} + \ln E_{.j} - \ln n$$
$$\Leftrightarrow \ln E_{.j} = \frac{1}{r}\sum_{i=1}^{r} \ln E_{ij} - \frac{1}{r}\sum_{i=1}^{r} \ln E_{i.} + \ln n \quad \text{..... (ii)}$$

- $\sum_{j=1}^{c} \ln E_{ij} = \sum_{j=1}^{c} \ln E_{i.} + \sum_{j=1}^{c} \ln E_{.j} - \sum_{j=1}^{c} \ln n$

$$\Leftrightarrow \sum_{j=1}^{c} \ln E_{ij} = c \ln E_{i.} + \sum_{j=1}^{c} \ln E_{.j} - c \ln n$$
$$\Leftrightarrow \frac{1}{c}\sum_{j=1}^{c} \ln E_{ij} = \ln E_{i.} + \frac{1}{c}\sum_{j=1}^{c} \ln E_{.j} - \ln n$$
$$\Leftrightarrow \ln E_{i.} = \frac{1}{c}\sum_{j=1}^{c} \ln E_{ij} - \frac{1}{c}\sum_{j=1}^{c} \ln E_{.j} + \ln n \quad \text{..... (iii)}$$

- $\sum_{i=1}^{r}\sum_{j=1}^{c} \ln E_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{c} \ln E_{i.} + \sum_{i=1}^{r}\sum_{j=1}^{c} \ln E_{.j} - \sum_{i=1}^{r}\sum_{j=1}^{c} \ln n$

$$\Leftrightarrow \sum_{i=1}^{r}\sum_{j=1}^{c} \ln E_{ij} = c\sum_{i=1}^{r} \ln E_{i.} + r\sum_{j=1}^{c} \ln E_{.j} - r\times c \ln n$$
$$\Leftrightarrow \frac{1}{r\times c}\sum_{i=1}^{r}\sum_{j=1}^{c} \ln E_{ij} = \frac{1}{r}\sum_{i=1}^{r} \ln E_{i.} + \frac{1}{c}\sum_{j=1}^{c} \ln E_{.j} - \ln n \quad \text{..... (iv)}$$

substituindo (ii) e (iii) em (i), obtive

$$\ln E_{ij} = \ln E_{i.} + \ln E_{.j} - \ln n$$
$$\Leftrightarrow \ln E_{ij} = \left(\frac{1}{c}\sum_{j=1}^{c} \ln E_{ij} - \frac{1}{c}\sum_{j=1}^{c} \ln E_{.j} + \ln n\right) + \left(\frac{1}{r}\sum_{i=1}^{r} \ln E_{ij} - \frac{1}{r}\sum_{i=1}^{r} \ln E_{i.} + \ln n\right) - \ln n$$
$$\Leftrightarrow \ln E_{ij} = \frac{1}{c}\sum_{j=1}^{c} \ln E_{ij} + \frac{1}{r}\sum_{i=1}^{r} \ln E_{ij} - \frac{1}{c}\sum_{j=1}^{c} \ln E_{.j} - \frac{1}{r}\sum_{i=1}^{r} \ln E_{i.} + \ln n$$
$$\Leftrightarrow \ln E_{ij} = \frac{1}{c}\sum_{j=1}^{c} \ln E_{ij} + \frac{1}{r}\sum_{i=1}^{r} \ln E_{ij} - \underbrace{\left(\frac{1}{r}\sum_{i=1}^{r} \ln E_{i.} + \frac{1}{c}\sum_{j=1}^{c} \ln E_{.j} - \ln n\right)}_{(iv)}$$
$$\Leftrightarrow \ln E_{ij} = \frac{1}{c}\sum_{j=1}^{c} \ln E_{ij} + \frac{1}{r}\sum_{i=1}^{r} \ln E_{ij} - \left(\frac{1}{r\times c}\sum_{i=1}^{r}\sum_{j=1}^{c} \ln E_{ij}\right)$$

Fazendo

$$\bar{n}_{i.} = \frac{1}{c}\sum_{j=1}^{c} \ln E_{ij}$$

$$\bar{n}_{.j} = \frac{1}{r}\sum_{i=1}^{r} \ln E_{ij}$$

$$\bar{n}_{..} = \frac{1}{r \times c}\sum_{i=1}^{r}\sum_{j=1}^{c} \ln E_{ij}$$

A equação anterior é equivalente a

$$\ln E_{ij} = \bar{n}_{i.} + \bar{n}_{.j} - \bar{n}_{..}$$

$$\Leftrightarrow \ln E_{ij} = \underbrace{(\bar{n}_{i.} - \bar{n}_{..})}_{\lambda_i^A} + \underbrace{(\bar{n}_{.j} - \bar{n}_{..})}_{\lambda_j^B} + \underbrace{\bar{n}_{..}}_{\mu}$$

Como o número de parâmetros é respetivamente 1, $r-1$ e $c-1$, logo, o número total dos parâmetros independentes é $r + c - 1$.

b) Modelo saturado ou modelo completo

Se não houver independência, o modelo anterior (*) torna-se inadequado, sendo necessário introduzir um termo representativo da interação entre as variáveis, seja λ_{ij}^{AB}, e o modelo fica

$$\ln E_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

Neste caso, $E_{ij} \neq \frac{O_{i.} \times O_{.j}}{n}$, pois H_0 de independência não é verdadeira

$$E_{ij} = n \times P(A_i \cap B_j) = n \times \frac{O_{ij}}{n} = O_{ij}$$

Com as restrições

$$\sum_{i=1}^{r} \lambda_i^A = 0,\ \sum_{j=1}^{c} \lambda_j^B = 0 \text{ e } \sum_{i=1}^{r} \lambda_{ij}^{AB} = \sum_{j=1}^{c} \lambda_{ij}^{AB} = 0$$

os parâmetros do modelo são estimados de forma equivalente aos parâmetros do modelo (*) substituindo E_{ij} por O_{ij} e adicionalmente $\hat{\lambda}_{ij}^{AB} = \ln O_{ij} - (\mu + \lambda_i^A + \lambda_j^B)$.

O número total de parâmetros independentes $1 + (r-1) + (c-1) + (r-1)\times(c-1) = r \times c$.

O modelo de independência e o modelo saturado não os únicos modelos possíveis.

- O modelo que inclui pelo menos todos os parâmetros relativos aos efeitos principais de cada uma das variáveis, chama-se *modelo abrangentes.*
- O modelo que não inclui pelo menos um parâmetro do efeito principal, chama-se *modelo não-abrangentes (noncomprehensive).*

Neste caso, estão a considerar os modelos log-lineares *abrangentes* que é resumido na tabela seguintes:

Tabela 7 : Modelos log-lineares abrangentes em tabela contingência bidimensional

Modelo	Parâmetros	Símbolo
Saturado	$\mu, \lambda_i^A, \lambda_j^B, \lambda_{ij}^{AB}$	(AB)
Independência	$\mu, \lambda_i^A, \lambda_j^B$	(A, B)

Interpretação dos parâmetros do modelo (Bento Murteira & Antunes, 2012)

- Se $\lambda_i^A > 0\ (< 0)$, o efeito da linha-*i* é positivo (negativo), i.e., as frequências esperadas da linha- i tendem a ser superiores (inferiores) à média global (μ).
- Se $\lambda_j^B > 0\ (< 0)$, o efeito da coluna-*j* é positivo (negativo), i.e., as frequências esperadas da coluna- j tendem a ser superiores (inferiores) à média global (μ).
- Se $\lambda_{ij}^{AB} > 0\ (< 0)$, há uma associação positiva (negativa) entre A e B. Para $\lambda_{ij}^{AB} = 0$, não existe associação entre A e B.

Ajustamento de modelos log-lineares

Verificar qual dos modelos, o de independência ou o saturado, melhor se ajusta aos dados.

Hipóteses a testar

H_0: modelo de independência ($\lambda_{ij}^{AB} = 0$)

H_1: modelo saturado ($\lambda_{ij}^{AB} \neq 0$)

Passos para ajustar os modelos log-lineares:

- Aplicar o teste do Qui-quadrado (χ^2) ou o teste de razão de verosimilhança (G^2), supondo H_0 verdadeiro. Se H_0 for rejeitada, considera-se o modelo saturado. Caso contrário, considera-se o modelo de independência.
- Estimar os parâmetros do modelo a considerar.
- Interpretar os resultados obtidos.

5. Tabelas de contingência $r \times c \times l$

Se uma tabela tem $\boldsymbol{r}$ linhas, $\boldsymbol{c}$ colunas e $\boldsymbol{l}$ estratos diz-se que tem dimensão $r \times c \times l$, e designa-se por tabela *tridimensional.* Se tem dimensão superior, a tabela de contingência designa-se por tabela *multidimensional.*

Sendo A, B e C três variáveis com r, c e l categorias, respetivamente, a tabela de contingência resultantes da classificação dos n indivíduos de uma amostra.

Tabela 8 : Tabela de contingência tridimensional

	$\boldsymbol{C_1}$			...	$\boldsymbol{C_l}$			**Totais**
	$\boldsymbol{B_1}$	...	$\boldsymbol{B_c}$	...	$\boldsymbol{B_1}$	...	$\boldsymbol{B_c}$	
$\boldsymbol{A_1}$	O_{111}	...	O_{1c1}	...	O_{11l}	...	O_{1cl}	$\boldsymbol{O_{i..}}$
⋮	⋮			⋮			⋮	⋮
$\boldsymbol{A_r}$	O_{r11}	...	O_{rc1}	...	O_{r1l}	...	O_{rcl}	$\boldsymbol{O_{r..}}$
Totais	$\boldsymbol{O_{.1.}}$	...	$\boldsymbol{O_{.c.}}$	...	$\boldsymbol{O_{.1.}}$	...	$\boldsymbol{O_{.c.}}$	$\boldsymbol{n}$

Esta tabela tem r linhas, c colunas, l estratos e $r \times c \times l$ células. Cada célula contém uma frequência observada, O_{ijk}, representa o número de elemento da amostra que verificam simultaneamente as categorias A_i de A, B_j de B e C_k de C.

A dimensão da amostra

$$n = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} O_{ijk}\,, i = 1, \dots, r, \mathrm{j} = 1, \dots, c, k = 1, \dots, l$$

Os totais marginais desta tabela são seguintes:

- Total marginal de uma só variável:

$$O_{i..} = \sum_{j=1}^{c}\sum_{k=1}^{l} O_{ijk}\,;\; O_{.j.} = \sum_{i=1}^{r}\sum_{k=1}^{l} O_{ijk}\,;\; O_{..k} = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ijk}\,; i = 1, \dots, r;\; \mathrm{j} = 1, \dots, c;\; k = 1, \dots, l$$

- Total marginal de duas variáveis:

$$O_{ij.} = \sum_{k=1}^{l} O_{ijk}\,;\; O_{i.k} = \sum_{j=1}^{c} O_{ijk}\,;\; O_{.jk} = \sum_{i=1}^{r} O_{ijk}\,; i = 1, \dots, r;\; \mathrm{j} = 1, \dots, c;\; k = 1, \dots, l$$

5.1. Hipóteses de independência

Neste caso teremos testar mais do que uma hipótese, pois pode-nos interessar estudar a associação ou independência das três variáveis entre si, de duas variáveis relativamente ao terceiro ou ainda de uma variável relativamente a cada uma das outras duas (Leal, 1997).

a) Independência mútua: três variáveis são independentes, ou não estão associados entre si.

Hipóteses a testar

H_0**:** as três variáveis são independentes

H_1: $\sim H_0$

Supondo H_0 de independência mútua verdadeira,

$$E_{ijk} = n \times P\left(A_i \cap B_j \cap C_k\right) = n \times P(A_i) \times P\left(B_j\right) \times P(C_k) = n \times \frac{O_{i..}}{n} \times \frac{O_{.j.}}{n} \times \frac{O_{..k}}{n} = \frac{O_{i..} \times O_{.j.} \times O_{..k}}{n^2},\ \forall ijk.$$

b) Independência parcial, existem três hipóteses para testar independência parcial

- Independência parcial entre C e (AB)

$H_0^{(1)}$**:** a variável C independente das restantes (AB)

$H_1^{(1)}$: $\sim H_0^{(1)}$

Supondo $H_0^{(1)}$ verdadeira,

$$E_{ijk} = n \times P\left(A_i \cap B_j \cap C_k\right) = n \times P\left(\left(A_i \cap B_j\right) \cap C_k\right) = n \times P\left(A_i \cap B_j\right) \times P(C_k)$$

$$= n \times \frac{O_{ij.}}{n} \times \frac{O_{..k}}{n} = \frac{O_{ij.} \times O_{..k}}{n};\ \forall i, j, k$$

- Independência parcial entre B e (AC)

  $H_0^{(2)}$**:** a variável B independente das restantes (AC)

  $H_1^{(2)}: \sim H_0^{(2)}$

  Supondo $H_0^{(2)}$ verdadeira, $E_{ijk} = \frac{O_{i.k} \times O_{.j.}}{n}; \forall i, j, k$

- Independência parcial entre A e (BC)

  $H_0^{(3)}$**:** a variável A independente das restantes (BC)

  $H_1^{(3)}: \sim H_0^{(3)}$

  Supondo $H_0^{(3)}$ verdadeira, $E_{ijk} = \frac{O_{.jk} \times O_{i..}}{n}; \forall i, j, k$

c) Independência condicional, existem 3 hipótese para testar independência condicional.

- Independência condicional entre A e B dada C

  $H_0^{(1)}$**:** as variáveis A e B são condicionalmente independentes de C

  $H_1^{(1)}: \sim H_0^{(1)}$

  Supondo $H_0^{(1)}$ verdadeira,

  $$E_{ijk} = n \times P\left(A_i \cap B_j \cap C_k\right) = n \times P\left(A_i \cap B_j | C_k\right) \times P(C_k) = n \times P(A_i | C_k) \times P\left(B_j | C_k\right) \times P(C_k)$$

  $$= n \times \frac{P(A_i \cap C_k)}{P(C_k)} \times \frac{P\left(B_j \cap C_k\right)}{P(C_k)} \times P(C_k) = n \times \frac{\frac{O_{i.k}}{n}}{\frac{O_{..k}}{n}} \times \frac{O_{.jk}}{n} = \frac{O_{i.k} \times O_{.jk}}{O_{..k}}; \forall i, j, k.$$

- Independência condicional entre A e C dada B

  $H_0^{(2)}$**:** as variáveis A e C são independentes dada variável B

  $H_1^{(2)}: \sim H_0^{(2)}$

  Supondo $H_0^{(2)}$ verdadeira, $E_{ijk} = \frac{O_{ij.} \times O_{.jk}}{O_{.j.}}; \forall i, j, k$

- Independência condicional entre B e C dada A

  $H_0^{(3)}$**:** as variáveis B e C são independentes dada variável A

  $H_1^{(3)}: \sim H_0^{(3)}$

  Supondo $H_0^{(3)}$ verdadeira, $E_{ijk} = \frac{O_{ij.} \times O_{i.k}}{O_{i..}}; \forall i, j, k$

d) Associação 2 a 2

Hipóteses a testar

H_0**:** a associação entre duas variáveis não depende das categorias da terceira

$H_1: \sim H_0$

Não é possível escrever explicitamente E_{ijk} em função dos valores de $O_{...}$ e é necessário utilizar um método iterativo para estimar E_{ijk}.

Método iterativo: ajustamento proporcional iterativo

Este método permite obter as estimativas das frequências esperadas E_{ijk} em modelo log-lineares hierárquicos. Este método iterativo começa por usar quaisquer estimativas iniciais, $\left\{\hat{E}_{ijk}^{(0)}\right\}$, desde que satisfaçam o modelo ajustar. Multiplicando estes valores por fatores de escala apropriados, ajustam-se sucessivamente as estimativas iniciais de modo a que os seus valores coincidam com as frequências marginais que consistem um conjunto mínimo de informação que é "suficiente" para obter E_{ijk} no modelo.

Estatística de teste

Em qualquer dos casos anterior, a estatística de teste pode ser:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\frac{\left(O_{ijk}-E_{ijk}\right)^2}{E_{ijk}} \text{ ou } G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\left(O_{ijk}\ln\left(\frac{O_{ijk}}{E_{ijk}}\right)\right)$$

onde E_{ijk} representa a frequência esperada estimada sob a hipótese que se pretende a testar. Seja qual for a hipótese H_0, tem-se: $\chi^2 \sim \boldsymbol{\chi_v^2}$ e $G^2 \sim \boldsymbol{\chi_v^2}$, v é grau de liberdade.

Na tabela a seguir estão resumindo as estimativas de frequências esperadas e graus de liberdade para cada hipótese de independência.

Tabela 9 : Estimativas de E_{ijk} para cada hipótese independência

Hipótese de independência	E_{ijk}	v
Mutua	$\frac{O_{i..}\times O_{.j.}\times O_{..k}}{n^2}$	$rcl-r-c-l+2$
Parcial entre (A, B) e C	$\frac{O_{ij.}\times O_{..k}}{n}$	$(rc-1)(l-1)$
Parcial entre (A, C) e B	$\frac{O_{i.k}\times O_{.j.}}{n}$	$(rl-1)(c-1)$
Parcial entre (B, C) e A	$\frac{O_{.jk}\times O_{i..}}{n}$	$(cl-1)(r-1)$
Condicional entre (A, B) dada C	$\frac{O_{i.k}\times O_{.jk}}{O_{..k}}$	$l(r-1)(c-1)$
Condicional entre (A, C) dada B	$\frac{O_{ij.}\times O_{.jk}}{O_{.j.}}$	$c(r-1)(l-1)$
Condicional entre (B, C) dada A	$\frac{O_{ij.}\times O_{i.k}}{O_{i..}}$	$r(c-1)(l-1)$
Associação 2 a 2	Método iterativo	$(r-1)(c-1)(l-1)$

5.2. Modelos log-lineares

À semelhança do efetuado para tabelas bidimensionais, nas tabelas tridimensionais devem ser considerar as seguintes notações:

$$\bar{n}_{...} = \frac{1}{r \times c \times l} \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{l} \ln E_{ijk}$$

$$\bar{n}_{i..} = \frac{1}{j \times k} \sum_{j=1}^{c} \sum_{k=1}^{l} \ln E_{ijk} \quad ; \quad \bar{n}_{.j.} = \frac{1}{i \times k} \sum_{i=1}^{r} \sum_{k=1}^{l} \ln E_{ijk} \quad ; \quad \bar{n}_{..k} = \frac{1}{i \times j} \sum_{i=1}^{r} \sum_{j=1}^{c} \ln E_{ijk}$$

$$\bar{n}_{ij.} = \frac{1}{k} \sum_{k=1}^{l} \ln E_{ijk} \quad ; \quad \bar{n}_{.jk} = \frac{1}{i} \sum_{i=1}^{r} \ln E_{ijk} \quad ; \quad \bar{n}_{i.k} = \frac{1}{j} \sum_{j=1}^{c} \ln E_{ijk}$$

e assim os parâmetros do modelo são estimados pelo:

$\mu = \bar{n}_{...}$; efeito médio global

$\lambda_i^A = \bar{n}_{i..} - \mu$; efeito da variável A

$\lambda_j^B = \bar{n}_{.j.} - \mu$; efeito da variável B

$\lambda_k^C = \bar{n}_{..k} - \mu$; efeito da variável B

$\lambda_{ij}^{AB} = \bar{n}_{ij.} - \bar{n}_{i..} - \bar{n}_{.j.} + \mu$; efeito da intersecção entre A e B

$\lambda_{ik}^{AC} = \bar{n}_{i.k} - \bar{n}_{i..} - \bar{n}_{..k} + \mu$; efeito da intersecção entre A e C

$\lambda_{jk}^{BC} = \bar{n}_{.jk} - \bar{n}_{.j.} - \bar{n}_{..k} + \mu$; efeito da intersecção entre B e C

$\lambda_{ijk}^{ABC} = \ln E_{ijk} - \bar{n}_{ij.} - \bar{n}_{i.k} - \bar{n}_{.jk} - \bar{n}_{i..} - \bar{n}_{.j.} - \bar{n}_{..k} - \mu$; efeito da intersecção entre A, B e C ou intersecção da 2ª ordem

Restrição adicional:

$$\sum_{i=1}^{r} \lambda_i^A = \sum_{j=1}^{c} \lambda_j^B = \sum_{k=1}^{l} \lambda_k^C = \sum_{i=1}^{r} \lambda_{ij}^{AB} = \cdots = \sum_{j=1}^{c} \lambda_{ijk}^{ABC} = \sum_{k=1}^{l} \lambda_{ijk}^{ABC} = 0$$

a) Modelo saturado

Tendo em conta o modelo saturado para tabela bidimensionais, obtém-se o seguinte modelo de saturado para tabela tridimensional:

$$\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

Neste caso, $E_{ij} = n \times P\left(A_i \cap B_j \cap C_k\right) = n \times \frac{O_{ijk}}{n} = O_{ijk}$

O número total de parâmetros independentes do modelo é igual a:

$1 + (r-1) + (c-1) + (l-1) + (r-1)(c-1) + (r-1)(l-1) + (c-1)(l-1) + (r-1)(c-1)(l-1) = rcl$

b) Modelo independência

Vamos considerar apenas" modelo hierárquicos", sabendo que: um modelo diz-se **hierárquico** se, ao incluir um termo representativo de um efeito de determinada ordem, envolvendo um conjunto de variáveis S, também inclui todos os termos que representam efeitos de ordens inferiores envolvendo qualquer subconjunto de S (por exemplo, se o modelo incluir o termo λ_{12}, também deverá incluir os termos λ_1 e λ_2). (Leal, 1997)

Cada um dos modelos especifica um dos tipos de independência a que nos referimos, pelo que a sua designação coincide com a atribuída a esses tipos de independência.

i) Modelo de associação 2 a 2

$$\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

ii) Modelo de independência condicional

- $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} \longrightarrow$ independência condicional entre A e B dada C
- $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} \longrightarrow$ independência condicional entre A e C dada B
- $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} \longrightarrow$ independência condicional entre B e C dada A

iii) Modelos de independência parcial

- $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} \longrightarrow$ independência parcial entre (AB) e C
- $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} \longrightarrow$ independência parcial entre (AC) e B
- $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{jk}^{BC} \longrightarrow$ independência parcial entre (BC) e A

iv) Modelo independência mútua

$$\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C$$

Na tabela a seguir resumir-se os modelos de log-lineares em tabelas de contingência tridimensional.

Tabela 10 : Modelos log-lineares abrangentes em tabela contingência tridimensional

Modelo		**Parâmetros**	**Símbolo**
Saturado/completo		$\mu, \lambda_i^A, \lambda_j^B, \lambda_k^C, \lambda_{ij}^{AB}, \lambda_{ik}^{AC}, \lambda_{jk}^{BC}, \lambda_{ijk}^{ABC}$	(ABC)
Associação 2 a 2		$\mu, \lambda_i^A, \lambda_j^B, \lambda_k^C, \lambda_{ij}^{AB}, \lambda_{ik}^{AC}, \lambda_{jk}^{BC}$	(AB, AC, BC)
Independência condicional	entre B e C, dado A	$\mu, \lambda_i^A, \lambda_j^B, \lambda_k^C, \lambda_{ik}^{AC}, \lambda_{jk}^{BC}$	(AB, AC)
	entre A e C, dado B	$\mu, \lambda_i^A, \lambda_j^B, \lambda_k^C, \lambda_{ij}^{AB}, \lambda_{jk}^{BC}$	(AB, BC)
	entre A e B, dado C	$\mu, \lambda_i^A, \lambda_j^B, \lambda_k^C, \lambda_{ij}^{AB}, \lambda_{ik}^{AC}$	(AC, BC)
Independência parcial	entre (B, C) e A	$\mu, \lambda_i^A, \lambda_j^B, \lambda_k^C, \lambda_{ij}^{CB}$	(BC, A)
	entre (A, C) e B	$\mu, \lambda_i^A, \lambda_j^B, \lambda_k^C, \lambda_{ik}^{AC}$	(AC, B)
	entre (A, B) e C	$\mu, \lambda_i^A, \lambda_j^B, \lambda_k^C, \lambda_{ij}^{AB}$	(AB, C)
Independência mútua		$\mu, \lambda_i^A, \lambda_j^B, \lambda_k^C$	(A, B, C)

Ajustamento do modelo

O objetivo final da análise log-linear consiste em determinar o modelo que, com o menor número possível de termos, descreve satisfatoriamente os dados.

Precisamos de averiguar quais os termos do modelo saturado que se podem eliminar, ou seja, que se podem considerar com não sendo significativamente não nulos. Para tal, poderemos recorrer a teste estatísticos: ajustar um modelo a uma tabela corresponde a testar a hipótese de independência especificada por esse modelo. Para a testar o ajustamento global dos modelos log-lineares usa-se a estatística qui-quadrado (χ^2) ou a estatística de razão de verosimilhança (G^2) dados por

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\frac{\left(O_{ijk} - E_{ijk}\right)^2}{E_{ijk}} \qquad \text{e} \qquad G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\left(O_{ijk}\ln\frac{O_{ijk}}{E_{ijk}}\right).$$

Quando o tamanho de amostra é elevado χ^2 e G^2 têm aproximadamente uma distribuição Qui-quadrado com v graus de liberdade (onde v dependendo do modelo considerado).

Nas tabelas a seguir estão a resumir as informações mínimas necessárias para estimar as frequências esperadas e estimativas de frequências esperadas de cada modelo.

Tabela 11 : Informação mínimas e estimativas de E_{ijk} para tabela tridimensional

Modelo		**Símbolo**	**Informação***	$\widehat{E}_{ijk}$
Saturado/completo		(ABC)	$\{O_{ijk}\}$	O_{ijk}
Associação 2 a 2		(AB, AC, BC)	$\{O_{ij.}\}, \{O_{i.k}\}, \{O_{.jk}\}$	Método iterativo
Independência condicional	entre B e C, dado A	(AB, AC)	$\{O_{ij.}\}, \{O_{i.k}\}$	$\frac{O_{ij.} \times O_{i.k}}{O_{i..}}$
	entre A e C, dado B	(AB, BC)	$\{O_{ij.}\}, \{O_{.jk}\}$	$\frac{O_{ij.} \times O_{.jk}}{O_{.j.}}$
	entre A e B, dado C	(AC, BC)	$\{O_{i.k}\}, \{O_{.jk}\}$	$\frac{O_{i.k} \times O_{.jk}}{O_{..k}}$
Independência parcial	entre (B, C) e A	(BC, A)	$\{O_{.jk}\}, \{O_{i..}\}$	$\frac{O_{.jk} \times O_{i..}}{n}$
	entre (A, C) e B	(AC, B)	$\{O_{i.k}\}, \{O_{.j.}\}$	$\frac{O_{i.k} \times O_{.j.}}{n}$
	entre (A, B) e C	(AB, C)	$\{O_{ij.}\}, \{O_{..k}\}$	$\frac{O_{ij.} \times O_{..k}}{n}$
Independência mútua		(A, B, C)	$\{O_{i..}\}, \{O_{.j.}\}, \{O_{..k}\}$	$\frac{O_{i..} \times O_{.j.} \times O_{..k}}{n^2}$

Seleção de modelos

Considere-se, dois modelos $\boldsymbol{M_a}$ e $\boldsymbol{M_b}$ com $\boldsymbol{v_a}$ e $\boldsymbol{v_b}$ respetivamente os graus de liberdade dos modelos considerado. Sendo $\boldsymbol{M_a}$ um caso particular de $\boldsymbol{M_b}$, então $\boldsymbol{M_a}$ e $\boldsymbol{M_b}$ dizem-se modelos "encaixados". A hipótese a testar neste caso é que os modelos $\boldsymbol{M_a}$ e $\boldsymbol{M_b}$ são igualmente bons (ou seja, de que os termos que estão em $\boldsymbol{M_a}$ e não estão em $\boldsymbol{M_b}$ são nulos).

Se decidir rejeitar esta hipótese, conclui-se que os termos que estão em $\boldsymbol{M_a}$ e não estão em $\boldsymbol{M_b}$ são significativamente não nulos, devendo, portanto, ser concluídos no modelo final. No caso contrário, conclui-se que estes termos devem ser eliminados.

A comparação deste tipo de modelos pode ser feita mediante o cálculo da estatística:

$$G^2(M_a|M_b) = G^2(M_a) - G^2(M_b) \sim \chi^2_{(|\boldsymbol{v_b}-\boldsymbol{v_a}|)}$$

Decisão: rejeita-se H_0 (em favor de H_1) quando $G^2(M_a|M_b) \geq \chi^2_{\alpha;(|\boldsymbol{v_b}-\boldsymbol{v_a}|)}$. Se $G^2(M_a|M_b) < \chi^2_{\alpha;(|\boldsymbol{v_b}-\boldsymbol{v_a}|)}$ então não se rejeita H_0.

Passos para ajustar os modelos log-lineares tabela tridimensional:

- Aplicar o teste do Qui-quadrado (χ^2) ou o teste de razão de verosimilhança (G^2), supondo H_0 verdadeiro. Se H_0 for rejeitada, considera-se o modelo saturado. Caso contrário, considera-se o modelo de independência, e neste caso, é necessário de fazer seleção de modelos encaixados.
- Estimar os parâmetros do modelo a considerar.
- Interpretar os resultados obtidos.

Capítulo 3 Regressão linear e análise de variância (ANOVA)

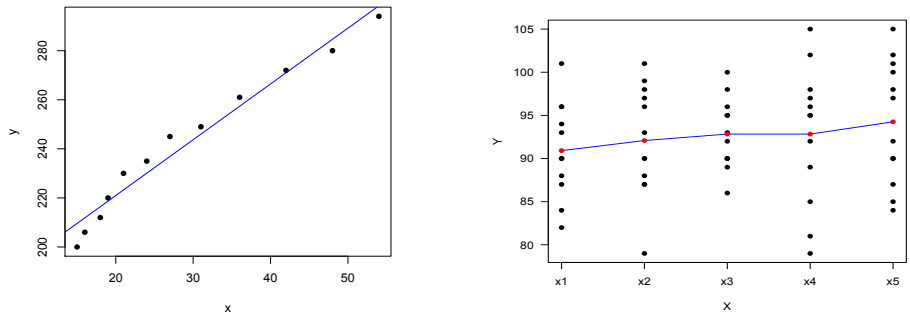
Neste capítulo, desenvolvem-se três temas. O primeiro inclui os conceitos de correlação e regressão, e desenvolve-se sobre a diagrama de dispersão e coeficiente de correlação. O segundo tema inclui o modelo de regressão linear (simples e múltiplo), incluindo a definição do modelo propriamente dito, estimação e inferência sobre os parâmetros, significado e avaliação da qualidade da regressão, validação de pressupostos desta análise e previsão de valores. A regressão linear múltipla inclui ainda, métodos de diagnostico de ponto de influentes, avaliação de colinearidade e métodos de seleção de variáveis. O terceiro tema inclui a análise de variância (ANOVA), desenvolvendo-se sobre ANOVA com um fator de efeitos fixos ou um fator de efeitos aleatórios, validação dos pressupostos desta análise e ANOVA não paramétrica com um factor.

1. Introdução

A análise de regressão linear visa modelar uma variável resposta numérica (quantitativa), à custa de uma ou mais variáveis preditoras, igualmente numéricas (quantitativo).

Mas na ANOVA é modelar uma variável resposta numérica (quantitativo) pode depender de variáveis categóricas (qualitativas), ou seja, de um ou mais fator.

Gráfico 7 : Diagrama de dispersão (Regressão) e gráfico de médias (ANOVA)



Regressão ANOVA

2. Correlação e regressão

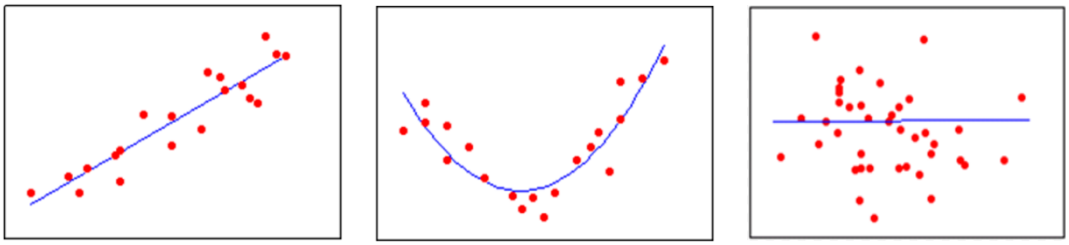
Para analisar o relacionamento entre duas variáveis (uma variável independente e uma variável dependente) é possível construir um modelo matemático que melhor representa este relacionamento, e que é obtido por análise de regressão. Alternativamente, se o objetivo é simplesmente medir o grau ou força do relacionamento entre as variáveis pode ser realizada uma análise de correlação.

Um dos métodos mais usados para estudar o tipo relacionamento é ***diagrama de dispersão***. É um gráfico onde cada ponto representa um par de valores observados (x_i, y_i) correspondentes, respetivamente aos valores das variáveis independentes e dependentes. O diagrama de dispersão tem uma dupla função:

- Ajuda a determinar se existe alguma relação entre as variáveis e
- Permite identificar qual é equação mais apropriadas para descrever essa relação.

A seguir estão representados os diagramas de dispersão de alguns tipos de relações.

Gráfico 8 : Diagrama de dispersão



Relação linear positiva | Relação não linear | Relação nula

Existem vários tipos de relações que podem estabelecer, mas apenas vamos considerar relação de tipo linear, $y = ax + b$. Quando o diagrama de dispersão indica uma tendência para uma relação linear, então os pontos encontram-se bem ajustados pela uma reta $y = ax + b$.

Se pretende relacionar duas variáveis, ou seja, um conjunto de pares de observações $(x_i, y_i), i = 1, \dots, n$, uma medida habitual do grau de relacionamento é o coeficiente de correlação de Pearson, definido por (Mello, 2014):

$$r = \frac{cov(x,y)}{\sqrt{var(x).var(y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 . \sum_{i=1}^{n} y_i^2 - n\bar{y}^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Em geral quando relacionamos duas variáveis aleatórios não encontramos uma relação linear perfeita do tipo $y = ax + b$. Como consequência o valor de r que se obtém pode ser qualquer valor no intervalo $[-1, 1]$ estando próximo dos extremos quando a relação entre as duas variáveis é forte e próxima da linear. Quando as variáveis são independentes então r será próxima de 0.

3. Regressão linear simples

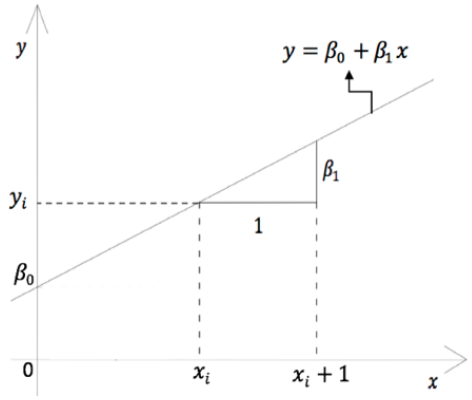
3.1. Modelo de regressão linear simples

No modelo de regressão linear simples queremos relacionar duas variáveis de um modelo linear, ou seja, através da equação de uma reta $y = \beta_0 + \beta_1 x$.

Este modelo tenta explicar a variação da variável y à custa do comportamento da variável x através de um modelo de regressão de y sobre x. Assim, diz-se que a variável x é a "variável explicativa ou regressora" (não é aleatória) e que y é a "variável explicada ou resposta" (é aleatória). O objetivo final consiste em estimar o valor da variável dependente em função da variável independente, ou seja, consiste em fazer uma previsão de valores futuros da variável dependentes.

Considere-se o conjunto de dados observados $(x_i, y_i), i = 1, \dots, n$. Neste modelo, em termos de análise de regressão, que *a única variável que pode ter algum erro associada é a variável* y_i, admitindo que a variável x_i é conhecida sem qualquer erro. Assim, a relação entre as variáveis x_i e y_i não pode ser descrita por um modelo matemático da forma $y_i = \beta_0 + \beta_1 x_i$, mas sim por um modelo do tipo $\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$, onde β_0 e β_1 são os parâmetros da regressão (β_0 é coeficiente constante ou intercepto, ou seja, o ponto de intersecção da reta com eixo yy, isto é quando $x = 0$ e β_1 é o β_1 representa a inclinação (declive) da reta regressora, expressando a taxa de mudança em y, ou seja, indica a mudança na média da distribuição de probabilidade de y para um aumento de uma unidade na variável x) e ε_i designada por erro**.** O erro ou resíduo *aleatório* descreve o afastamento na vertical dos pontos em relação à reta de equação $y_i = \beta_0 + \beta_1 x_i$.

Gráfico 9 : Reta de regressão $\boldsymbol{y = \beta_0 + \beta_1 x}$



Verifica-se que nem todos os pontos se encontram sobre a reta de regressão e essa diferença é o erro. Mas supõe-se que a média desses erros tende a anular-se, ou seja, $E[\varepsilon_i] = 0$. Isso é mesmo que dizer que a variável y_i pensada como função de x_i tem média $\beta_0 + \beta_1 x_i$, ou seja, se observamos vários valores de y_i para a mesma abcissa x_i eles devem dispor-se verticalmente em torno do ponto sobre a reta e a sua média (valor esperado) deve ser exatamente a ordenada do ponto da reta. Analiticamente isso significa que existe uma relação linear entre o valor esperado de y_i e o valor do regressor que lhe corresponde, x_i,

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i$$

Assim, uma equação de regressão permite estimar valores de y, com base em valores conhecidos através de

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

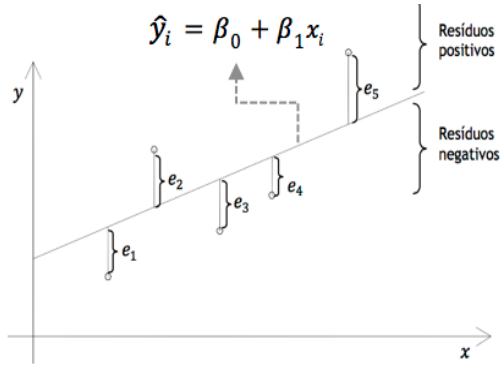
Assim, o modelo de regressão linear, $\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$ pode ser escrever na forma:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Leftrightarrow y_i = \hat{y}_i + \varepsilon_i$$

E o erro escreve-se

$$\varepsilon_i = y_i - \hat{y}_i \Leftrightarrow \varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

Gráfico 10 : Conjunto dos pontos (x_i, y_i), reta de regressão e os erros



3.2. Estimação e inferência sobre os parâmetros

a) Estimação dos parâmetros

Sendo o objetivo definir um modelo em que os erros cometidos sejam mais pequenos possíveis, a determinação dos parâmetros do modelo de regressão β_0 e β_1, é feita tal que seja mínima a soma dos quadrados erros, S. Portanto, as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ devem minimizar a soma dos quadrados dos erros.

$$\min_{\hat{\beta}_0, \hat{\beta}_1} S \text{ onde } S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Para tornar S mínima, é necessário derivar a função em ordem β_0 e β_1, igualar a zero e verificar o sinal da segunda derivada (é positiva). Assim,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Onde, S_{xy} é a variância conjunta amostral entre x e y, S_{xx} é a variância amostral de x e $\bar{y}\ e\ \bar{x}$ representam as médias amostral de y e x, respetivamente.

Demonstração:

$$\begin{cases} \frac{\partial}{\partial \beta_0}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2=0 \\ \frac{\partial}{\partial \beta_1}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2=0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n}-2(y_i-\beta_0-\beta_1 x_i)=0 \\ \sum_{i=1}^{n}-2x_i(y_i-\beta_0-\beta_1 x_i)=0 \end{cases} \Leftrightarrow \begin{cases} -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0 \\ -2\sum_{i=1}^{n}(x_i y_i-\beta_0 x_i-\beta_1 x_i^2)=0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0\sum_{i=1}^{n} x_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1\sum_{i=1}^{n} x_i}{n} \\ -\hat{\beta}_1\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} y_i - \hat{\beta}_1\sum_{i=1}^{n} x_i\right)\times\sum_{i=1}^{n} x_i}{n} = -\sum_{i=1}^{n} x_i y_i \end{cases}$$

$$\Leftrightarrow \begin{cases} -\hat{\beta}_1\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\sum_{i=1}^{n} y_i - \hat{\beta}_1\sum_{i=1}^{n} x_i\sum_{i=1}^{n} x_i\right)}{n} = -\sum_{i=1}^{n} x_i y_i \end{cases}$$

$$\Leftrightarrow \begin{cases} \hat{\beta}_1\left(-\sum_{i=1}^{n} x_i^2 + \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right) = -\sum_{i=1}^{n} x_i y_i + \frac{1}{n}\sum_{i=1}^{n} x_i\sum_{i=1}^{n} y_i \end{cases} \Leftrightarrow \begin{cases} \hat{\beta}_1 = \frac{-\sum_{i=1}^{n} x_i y_i + \frac{1}{n}\sum_{i=1}^{n} x_i\sum_{i=1}^{n} y_i}{-\sum_{i=1}^{n} x_i^2 + \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} \end{cases}$$

$$\Leftrightarrow \begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} \end{cases} \Leftrightarrow \begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i\, n\frac{1}{n}\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} \end{cases}$$

$$\Leftrightarrow \begin{cases} \boldsymbol{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}} \\ \boldsymbol{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}} \end{cases}$$

Estimação dos parâmetros do modelo (Notação matricial)

O modelo de regressão linear simples $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n$, pode ser escrito na forma matricial

$$\begin{matrix}(x_i, y_i) \\ i = 1, \dots, n\end{matrix} \longrightarrow \begin{cases} y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_n + \varepsilon_n \end{cases} \Leftrightarrow \underbrace{\begin{bmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{bmatrix}}_{n\times 1} = \underbrace{\begin{bmatrix} 1 & x_1 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}}_{n\times 2} \times \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{2\times 1} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}}_{n\times 1} \Leftrightarrow \boldsymbol{Y = X.\beta + \varepsilon}$$

A soma dos quadrados dos erros é dada por,

$$\mathrm{S} = \sum_{i=1}^{n} \boldsymbol{\varepsilon_i^2} = \boldsymbol{\varepsilon_i^T \varepsilon_i} = (\boldsymbol{Y - X\beta})^T(\boldsymbol{Y - X\beta}) = \boldsymbol{Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta}$$
$$= \boldsymbol{Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta} \quad \text{..... (**)}$$

O valor mínimo para S é obtido da seguinte maneira:

$$\frac{\partial}{\partial\beta}S = 0 \Leftrightarrow \frac{\partial}{\partial\beta}\boldsymbol{Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta} = 0$$
$$\Leftrightarrow \boldsymbol{-2X^T Y + 2X^T X\beta} = 0 \Leftrightarrow \boldsymbol{X^T Y = X^T X\beta}$$

Portanto, o estimador é

$$\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Assim, desde que $\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$, então a expressão (**) pode ser simplificar da seguinte forma:

$$\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i^2 = \boldsymbol{Y}^T\boldsymbol{Y} - 2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} + \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{Y}^T\boldsymbol{Y} - 2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} + \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y}$$

Hipóteses básicas do M.R.L.S.

- Linearidade entre x e y
- $E[\varepsilon_i|x] = 0$ e $var[\varepsilon_i|x] = \sigma^2 > 0$, (independente de x) ⟶ homocedasticidade condicionada
- $\varepsilon_i \sim N(0, \sigma^2)$ ⟶ normalmente distribuídas (e independentes)

Ao assumir que os erros têm distribuição normal concluímos que também as observações y_i vão ter distribuição normal já que $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, e soma de uma normal com uma constante tem distribuição normal. Assim para cada abcissa x_i

$$y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), i = 1, \dots, n.$$ (Hall et al., 2011)

estimação de σ^2

para estimar σ^2 precisar de conhecer ε_i, uma vez que a estimar σ^2 é através da variância amostral dos valores $\varepsilon_i, i = 1, \dots, n$. Assim,

$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(\varepsilon_i - 0)^2}{n-2} = \frac{\sum_{i=1}^{n}\varepsilon_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$ (Bento Murteira, Ribeiro, Silva, Pimenta, & Pimenta, n.d.)

b) Inferência sobre os parâmetros

Para realizar inferência sobre os parâmetros, é necessário conhecer a distribuição amostral dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$. Considere o modelo de regressão linear $Y = X.\beta + \varepsilon$, cujos estimadores de β é dado por $\hat{\beta} = [X^T X]^{-1} X^T Y$.

(i) Distribuição amostral dos parâmetros da regressão

Com base nos pressupostos do modelo de regressão linear simples podemos calcular a esperança e a variância dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$.

- Valor esperado

  Em forma matricial, $\widehat{\boldsymbol{\beta}} = [X^T X]^{-1} X^T Y = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \overbrace{[\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{X}}^{I}\boldsymbol{\beta} + [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}$

  Então conclui-se que

  $$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}$$

  Assim,

  $E(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = E(\boldsymbol{\beta} + [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}|\boldsymbol{X}) = E(\boldsymbol{\beta}) + E([\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{\beta} + [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\overbrace{E(\boldsymbol{\varepsilon}|\boldsymbol{X})}^{=0} = \boldsymbol{\beta}.$

- Variância/covariância

$$\Sigma_{\widehat{\boldsymbol{\beta}}} = var(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = var([\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{Y}|\boldsymbol{X}) = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T \overbrace{var(\boldsymbol{Y}|\boldsymbol{X})}^{=var(\boldsymbol{\varepsilon}|\boldsymbol{X})=\sigma^2} [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}$$

$$= \overbrace{[\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{X}}^{I} [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\sigma^2 = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\sigma^2$$

Assim, a matriz covariância dos parâmetros, é dada por:

$$\Sigma_{\widehat{\boldsymbol{\beta}}} = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\sigma^2 = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} & \frac{-\bar{x}}{s_{xx}} \\ \frac{-\bar{x}}{s_{xx}} & \frac{1}{s_{xx}} \end{bmatrix}\sigma^2 = \begin{bmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) \\ cov(\hat{\beta}_0, \hat{\beta}_1) & var(\hat{\beta}_1) \end{bmatrix}$$

Demonstração:

$$[\boldsymbol{X}^T\boldsymbol{X}] = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \Leftrightarrow [\boldsymbol{X}^T\boldsymbol{X}]^{-1} = \frac{1}{s_{xx}} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} = \frac{1}{s_{xx}} \times \frac{1}{n} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

$$= \frac{1}{s_{xx}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\frac{1}{n}\sum_{i=1}^{n} x_i \\ -\frac{1}{n}\sum_{i=1}^{n} x_i & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} & \frac{-\bar{x}}{s_{xx}} \\ \frac{-\bar{x}}{s_{xx}} & \frac{1}{s_{xx}} \end{bmatrix}$$

$var(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2 c_{ii}$ e $cov(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2 c_{ij};\ i \neq j$. Onde, c_{ii} é elementos da diagonal principal da $\Sigma_{\widehat{\boldsymbol{\beta}}}$ e c_{ij} é elementos que não sejam da diagonal principal.

Logo, a distribuição de cada parâmetro $\hat{\beta}_0$ e $\hat{\beta}_1$ é $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} \sim N(\boldsymbol{\beta}_i, \sigma^2 c_{i+1,i+1}),\ i = 0{,}1$.

(ii) Teste e intervalos de confiança para os parâmetros da regressão

De notar que ambos os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ são centrados e que a variância é proporcional à variâncias dos erros, σ^2. Isto significa que, tal como seria de esperar, se as observações estiverem sujeitas os erros elevados a incerteza com que estimamos a reta, quer quanto ao declive quer quanto à ordenada na origem, é grande.

No primeiro caso, vamos testar se a ordenada na origem é nula.

Hipóteses a testar são:

$H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 \neq 0$

A estatística de teste é

$T_0 = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} \sim t_{(n-2)}$, onde $S_{\hat{\beta}_0}$ é desvio padrão de $\hat{\beta}_0$

Decisão: rejeitar H_0 em favor de H_1 se $|T_{0obs}| > t_{(1-\alpha/2, n-2)}$. Caso contrário não rejeitar H_0.

No entanto, se não rejeitar H_0, podemos eliminar o parâmetro β_0 no modelo e passamos a ter menos um parâmetro desconhecido.

O Intervalo de confiança $(1-\alpha)\times 100\%$ para β_0 é $\left[\hat{\beta}_0 - t_{(1-\alpha/2,n-2)}S_{\hat{\beta}_0}\,;\hat{\beta}_0 + t_{(1-\alpha/2,n-2)}S_{\hat{\beta}_0}\right]$

No segundo caso, estamos a testar se de facto a variável y_i depende de x_i.

Hipóteses a testar são:

$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

A estatística de teste é

$T_1 = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \sim t_{(n-2)}$, onde $S_{\hat{\beta}_1}$ é desvio padrão de $\hat{\beta}_1$

Decisão: rejeitar H_0 em favor de H_1 se $|T_{1obs}| > t_{(1-\alpha/2,n-2)}$. Caso contrário não rejeitar H_0.

Se H_0 não foi rejeitada, ficamos com o modelo $y_i = \beta_0 + \varepsilon_i$ e não existe nenhuma relação entre y_i e x_i.

O Intervalos de confiança $(1-\alpha)$ para β_1 é $\left[\hat{\beta}_1 - t_{(1-\alpha/2,n-2)}S_{\hat{\beta}_1}\,;\hat{\beta}_1 + t_{(1-\alpha/2,n-2)}S_{\hat{\beta}_1}\right]$.

3.3. Significado e avaliação da qualidade da regressão

a) Significado estatístico do modelo

(i) Teste ao declive $\boldsymbol{\beta_1}$

Verificar se as variáveis independentes x_i contribuem significativamente com informação para explicar linearmente a variação da variável resposta y_i, como o que já referido anteriormente.

(ii) Teste ANOVA da regressão

É usual efetuar uma ANOVA no contexto da análise de regressão devendo ainda verificar se o modelo é ou não estatisticamente significativo e, em caso afirmativo, verificar se as variáveis x_i contribuem significativamente com informação para explicar y_i. Quanto mais isso for melhor será a previsão.

Vamos introduzir, de seguida, alguns conceitos e procedimentos básicos do método ANOVA.

Para analisar a significância, passemos à partição da soma dos quadrados através do modelo ajustado, isto é, a soma dos quadrados dos desvios das observações em relação à sua média, $\boldsymbol{SQ_T}$ é igual à soma dos quadrados dos desvios dos valores preditos (sobre a reta estimada) em relação à média, $\boldsymbol{SQ_R}$, com os quadrados dos desvios das observações em relação ao respetivo valor predito, $\boldsymbol{SQ_E}$.

$$SQ_T = SQ_R + SQ_E \Leftrightarrow \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Onde, SQ_T é a variabilidade total do conjunto de observações de y_i;

SQ_R é a variabilidade explicada pelo modelo;

SQ_E é a variabilidade explicada pelo modelo.

As somas de quadrados de desvios têm associado um número de graus de liberdade e à partição da variabilidade total SQ_T nos componentes SQ_R e SQ_E corresponde uma partição dos graus de liberdade. (Pedrosa & Gama, 2016).

Assim,

- SQ_R tem k graus de liberdade, onde k é o número de variáveis explicativas;
- SQ_E tem $n-p$ graus de liberdade, que resultam n resíduos $y_i - \hat{y}_i$ menos número de parâmetros estimados, onde $p = k + 1$.
- SQ_T tem $k + n - (k+1) = n - 1$ graus de liberdade.

O outro conceito importante é o de média de quadrados de desvios.

A média de quadrados de desvios é o quociente entre uma soma de quadrados de desvios e os correspondentes graus de liberdades (Pedrosa & Gama, 2016).

Na regressão há duas médias quadrados de desvios importantes: a média de quadrados dos desvios explicados pela regressão, $\boldsymbol{MQ_R}$, e a média de quadrados dos resíduos (erros ajustamento), $\boldsymbol{MQ_E}$.

$$MQ_R = \frac{SQ_R}{k} \text{ e } MQ_E = \frac{SQ_E}{n-p} = \sigma^2$$

Estatística de teste:

$$F = \frac{MQ_R}{MQ_E} \sim F_{k,n-p}$$

Neste caso, em modelo de regressão linear simples $k = 1$.

A região de rejeição definida através da distribuição $F_{k,n-p}$, para valores elevados da estatística de teste F, ou seja, rejeitar H_0 em favor de H_1 se $F_{obs} > F_{1-\alpha,1,n-2}$. Caso contrário não se rejeita H_0.

Se a regressão não tiver significado, a primeira parcela será fortemente reduzida e, então o declive da reta deve ser nulo, $\beta_1 = 0$.

A tabela da ANOVA correspondente é a seguinte:

Tabela 12 : Tabela ANOVA de modelo de regressão simples

Fonte de variação	$\boldsymbol{SQ}$	$\boldsymbol{g.l.}$	$\boldsymbol{MQ}$	$\boldsymbol{F_{obs}}$	$\boldsymbol{valor - P}$
Regressão (explicada)	SQ_R	1	$MQ_R = \frac{SQ_R}{1}$	$\frac{MQ_R}{MQ_E}$	$P(F > F_{obs})$
Erros (não explicada)	SQ_E	$n-2$	$MQ_E = \frac{SQ_E}{n-2}$		
Total	SQ_T	$n-1$			

b) Avaliação da qualidade do modelo

(i) Coeficiente de determinação

O coeficiente de determinação pode ser pensado como uma medida da quantidade de variabilidade explicada pelo modelo de regressão já que consiste na razão entre a soma dos quadrados desvios aos resíduos e a soma dos quadrados total. (Pedrosa & Gama, 2016)

$$R^2 = \frac{SQ_R}{SQ_T} = \frac{SQ_T - SQ_E}{SQ_T} = 1 - \frac{SQ_E}{SQ_T}, 0 \leq R^2 \leq 1$$

- Se $R^2 = 1 \Longrightarrow SQ_E = 0 \Leftrightarrow \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0 \Leftrightarrow y_i = \hat{y}_i, i = 1, \dots, n$
  $\Longrightarrow$ existe uma relação linear perfeito (quando todas as observações estão sobre a reta de regressão).
- Se $R^2 = 0 \Longrightarrow SQ_R = 0 \Leftrightarrow SQ_T = SQ_E$
  $\Longrightarrow$ não existe uma relação linear entre as variáveis (quando as observações estão a espalhar em forma aleatoriamente).

(ii) Coeficiente de determinação ajustado

É definido a partir de R^2 e ajustado com base na dimensão da amostra (n) e no número de parâmetros do modelo (p), para quantificar o grau do ajustamento do modelo. No caso do modelo de regressão simples, $p = 2$.

$$R_a^2 = 1 - \frac{\frac{SQ_E}{n-p}}{\frac{SQ_T}{n-1}}, 0 \leq R_a^2 \leq 1$$

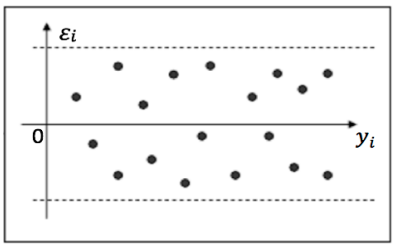
- $R_a^2 \cong 1$, o modelo é bastante adequado
- $R_a^2 \cong 0$, o modelo é pouco adequado

3.4. Validação dos pressupostos da regressão

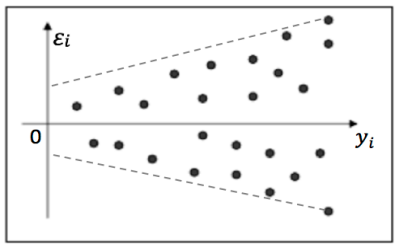
Uma forma de verificar que são validos os pressupostos do modelo consiste em efetuar uma análise dos resíduos ε_i, que deverão refletir as propriedades dos erros: serem normais, com variância constantes e independentes. Em particular, interessa verificar se os pressupostos dos erros são satisfeitos.

A análise dos resíduos através *do gráfico dos resíduos ε_i em função do valor preditos (valor ajustado) $\hat{y}_i$*, se os pontos do gráfico devem distribuir-se de forma aleatória em torno da reta que corresponde ao resíduo zero*,* formando uma mancha de largura uniforme. Dessa forma, será de esperar que os erros sejam "independentes", de "média nula" e de "variância constante". (Hall et al., 2011).

Gráfico 11 : Resíduos versus valores preditos



Homocedasticidade
(variância constante)



Heterocedasticidade
(variância não constante)

No caso de independência, para obter um resultado mais objetivo, aplicando o teste estatístico de independência, neste caso aplica o teste de **Durbin-Watson**. É um teste para detetar correlação entre resíduos sucessivos. Se os resíduos forem independentes, a magnitude de um resíduo não influencia a magnitude do resíduo seguinte.

Hipóteses a testar:

$H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

Sendo ρ é a correlação entre resíduos sucessivos

Estatística de teste: $dw = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2}$ onde dw é um valor tal que $0 \leq dw < 4$

Valores críticos: d_L e d_U (tabelados)

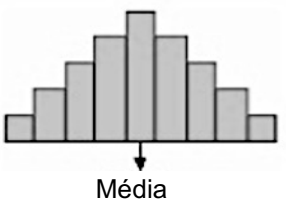
Decisão:

- $d_U \leq dw < 4 - d_U$, não se rejeita H_0
- $0 \leq dw < d_L$ ou $4 - d_L \leq dw < 4$, rejeita-se H_0
- $d_L \leq dw < d_U$ ou $4 - d_U \leq dw < 4 - d_L$, nada se pode concluir

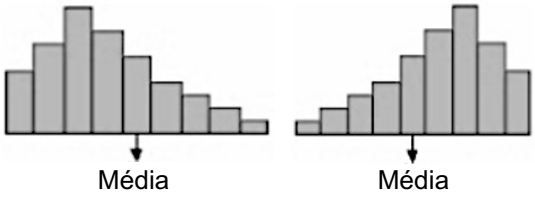
Para averiguar se os "erros têm distribuição Normal" pode ser visualizar graficamente através de *histograma* ou traçar *um QQ-plot*, e testar a hipótese de Normalidade.

- No *histograma* dos $\boldsymbol{\varepsilon_i}$, se os erros possuírem distribuição Normal: observa-se: concentração de valores em torno de um valor central; simétrica em torno do valor central; Frequência pequena de valores muito extremos.

Gráfico 12 : Histograma dos erros



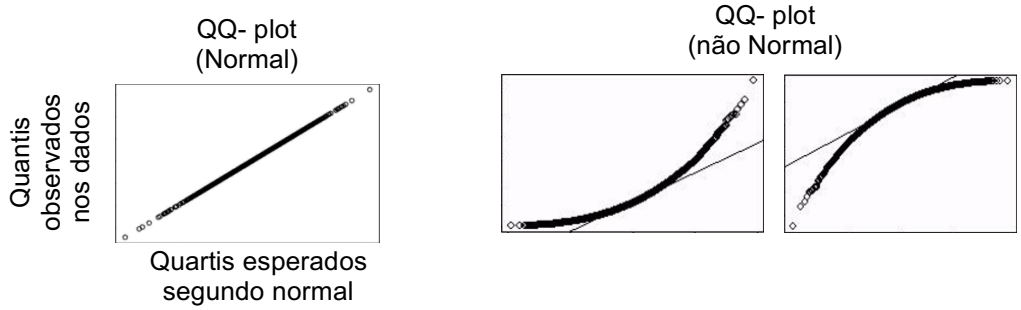
(a) Histograma (Normal)



(b) Histograma (não Normal)

- No “normal QQ-plot”, se os erros possuírem distribuição Normal, todos os pontos do gráfico devem posicionarem-se sobre a reta de referência.

Gráfico 13 : QQ-plot



- Teste de hipótese de Normalidade

Os métodos gráficos citados anteriormente têm a desvantagem de serem subjetivos, pois dependem de interpretação visual. Para um resultado mais objetivo, pode-se usar os testes de normalidade. Alguns deles, são: teste Kolmogorov-Smirnov (KS), Teste de normalidade de Lilliefors e Teste de normalidade de Shapiro-Wilk.

Teste de Kolmogorov-Smirnov (KS)

Consiste na comparação da função de distribuição acumulada (f.d.a.) dos valores observados e da função de distribuição acumulada teórica, de acordo com H_0, e na determinação do ponto de maior distância vertical entre as duas funções. Este teste é construído para qualquer distribuição e é válido para amostras de grande dimensão.

Assim, as hipóteses a testar são:

$H_0 : F(x) = F_0(x)$ vs. $H_1 : F(x) \neq F_0(x)$

Sendo $F(x)$ a função distribuição da variável aleatória continua e $F_0(x)$ é a função de distribuição acumulada teórica.

Estatística de teste:

$$D_n = \sup_x |\hat{F}_n(x) - F_0(x)|$$

Onde: $\hat{F}_n(x) = \frac{\sharp\, observações \leq x}{n}$ é f.d.a. empírica

$F_0(x) = \mathrm{P}(\mathrm{X} \leq x)$ é f.d.a. em H_0

Decisão: rejeita-se H_0 em favor de H_1 se “D_n for um valor muito alto”, concretamente se $D_n \geq d_{\alpha, n}$ (tabelado). Caso contrário não se rejeita H_0.

Teste de normalidade de Lilliefors

Este teste é uma adaptação do teste KS para o teste normalidade, e adequado para amostra de grande dimensão, $n \geq 30$. (Hall et al., 2011). Neste teste de Lilliefors usa a mesma estatística do Teste de KS, mas a tabela de valores críticos é a Tabela de Teste de Lilliefors, que é usada em vez da Tabela de KS.

Teste de normalidade de Shapiro-Wilk

É uma alternativa ao teste de KS para testar se a variável em estudo na amostra aleatória possui, ou não, distribuição normal e é adequado para amostras pequenas, $n < 30$. (Hall et al., 2011)

Hipóteses a testar:

$H_0 : X \sim N(\mu, \sigma^2)$ vs. $H_1 : \sim H_0$

Estatística de teste:

$W = \frac{b^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$,

onde $b = \begin{cases} \sum_{i=1}^{\frac{n}{2}} a_{n-i+1}\left(x_{(n-i+1)} - x_{(i)}\right) & se\ n\ é\ par \\ \sum_{i=1}^{\frac{n+1}{2}} a_{n-i+1}\left(x_{(n-i+1)} - x_{(i)}\right) & se\ n\ é\ impar \end{cases}$; em que a_{n-i+1} são tabelados

$\bar{x}$ é média da observações

*x_i *estão ordenados por ordem crescente*

Decisão: rejeitar H_0 em favor de H_1 se $W_{obs} < W_{\alpha,n}$ (tabelado), caso contrário não se rejeita H_0.

3.5. Previsão

Depois de termos introduzido os conceito básicos, vamos, agora, mostrar como usar o modelo de regressão linear para prever a resposta dada pelo modelo. (B Murteira, Ribeiro, Silva, & Pimenta, 2014) O problema da previsão procura dar resposta a dois tipos de questões:

a) Previsão em média

É estimação do valor esperado das observações do regressando associado a uma ou a várias combinações de valores assumidos pelos regressores. No caso previsão em média, pretende-se estimar o parâmetro

$$\theta = E(y_i | x_{i1} = c_1, x_{i2} = c_2, \dots, x_{ik} = c_k) = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k,$$

com k é número de regressores.

No caso de modelo regressão linear simples $k = 1$. Neste caso, o estimador de θ é

$$\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_1 c_1.$$

Fazendo $c = [1 \quad c_1]$ então o valor esperado e variância do estimador de θ são seguintes:

$$E(\hat{\theta}|x,c) = E(\hat{\beta}_0 + \hat{\beta}_1 c_1|x,c) = \beta_0 + \beta_1 c_1 = \theta$$

$$Var(\hat{\theta}|x,c) = Var(c\hat{\beta}|x,c) = c\ Cov(\hat{\beta}|x,c)c^T = \sigma^2 c(X^T X)^{-1}c^T \ldots\ldots\ldots\ldots (*)$$

A raiz quadrada de (*) é o **erro padrão da previsão em média,**

$$s_{\hat{\theta}} = \hat{\sigma}\sqrt{c(X^T X)^{-1}c^T}$$

e, assim, o I.C. de previsão para θ com confiança $(1-\alpha)$ é dado por

$$\left[\hat{\theta} - t_{1-\alpha/2}, s_{\hat{\theta}}\ ; \hat{\theta} + t_{1-\alpha/2, n-2} s_{\hat{\theta}}\right].$$

b) Previsão pontual (para valores isolados)

É estimação de valores observados pelo regressando em correspondência com uma ou várias combinações de valores assumidos pelos regressores.

Nalguns casos, especialmente com dados temporais, a previsão em média não tem interesse, devido à própria natureza dos dados. Com efeito, em muitas situações não tem sentido prever o comportamento médio do regressando, estando o investigador interessado em fazer previsão pontual, isto é, prever apenas um particular valor desta variável referido a outra período ou outra contexto. (Bento Murteira et al., n.d.)

Considere-se de novo que $x_{i1} = c_1, x_{i2} = c_2, \ldots, x_{ik} = c_k$, ou seja, neste caso $k = 1$, então

$$y_0 = \beta_0 + \beta_1 c_1 + \varepsilon_0$$

Enquanto que na previsão em média se pretendia estimar $E(y_0|x,c)$, na previsão pontual procura-se prever os valores assumidos por y_0.

Considere-se o previsor mínimo quadrado de y_0,

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1$$

e o erro de previsão,

$$\varepsilon = y_0 - \hat{y}_0$$

utilizando a variável aleatória ε, vão estudar-se as propriedades estatísticas do previsor. Como

$$E(\varepsilon|x,c) = E(y_0 - \hat{y}_0|x,c) = 0$$

a variância de ε, condicionada por X e c, é dada por

$$Var(\varepsilon|x,c) = \sigma^2\{1 + c(X^T X)^{-1}c^T\}$$

Assim, o **erro padrão da previsão** é dado por

$$s_\varepsilon = \hat{\sigma}\sqrt{1 + c(X^T X)^{-1}c^T}$$

Portanto, o I.C. de previsão para y_0 com confiança $(1-\alpha)$ é dado por

$$\left[\hat{y}_0 - t_{1-\alpha/2}, s_\varepsilon\ ; \hat{y}_0 + t_{1-\alpha/2, n-2} s_\varepsilon\right].$$

4. Regressão linear múltipla

4.1. Modelo de regressão linear múltipla

Se num modelo de regressão linear simples introduzirmos mais regressores passamos a ter um modelo de regressão linear múltipla. Neste caso estaremos a relacionar uma variável dependente y com mais do que uma variável independente.

A análise de um modelo de regressão linear múltipla é analoga à do modelo de regressão linear simples, sendo as analogias mais fáceis de identificar se usarmos uma notação matricial.

Assim, o modelo de regressão linear múltipla

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, i = 1, \dots, n, j = 1, \dots, k$$

ou,

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, i = 1, \dots, n$$

onde, i é o número de observações

j é o numero de regressores

x_{ij} é o valor de observação i na variável x.

$\beta_j, j = 0,1, \dots, k$ são os parâmetros da regressão

$p = k + 1$ é o número de parâmetros do modelo

Em notação matricial, deve ser escrito como

$$\begin{matrix}(x_{ij}, y_i) \\ i = 1, \dots, n \\ j = 1, .., k\end{matrix} \longrightarrow \begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \cdots \beta_k x_{1k} + \varepsilon_1 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_k x_{nk} + \varepsilon_n \end{cases} \Leftrightarrow \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{nk} & \dots & x_{nk} \end{bmatrix}}_{n \times p} \times \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{p \times 1} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{n \times 1}$$

$$\Leftrightarrow \boldsymbol{Y} = \boldsymbol{X}.\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

4.2. Estimação e inferência sobre os parâmetros

As estimativas dos p parâmetros da regressão ($\hat{\beta}_0,\ \hat{\beta}_1, \dots, \hat{\beta}_k$) são dadas pelas soluções da minimização da soma dos quadrados dos erros.

$$\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^T \boldsymbol{X}]^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

onde, $\boldsymbol{X}^T$ representa a matriz transposta de $\boldsymbol{X}$**.** Assim, a equação da superfície de regressão pode ser escrita como,

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

onde, $\widehat{\boldsymbol{Y}}$ é o vetor de valores preditos.

Os resíduos continuam a ser definidos como a diferença entre as observações e os valores preditos respetivos.

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$$

As somas dos quadrados dos erros são dadas por

$$SQ_E = \sum_{i=1}^{n} \boldsymbol{\varepsilon_i^2} = \boldsymbol{\varepsilon_i^T \varepsilon_i} = (\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta}) = \boldsymbol{Y^T Y} - \boldsymbol{\hat{\beta}^T X^T Y}$$

$$SQ_R = \boldsymbol{\hat{\beta}^T X^T Y} - n\bar{y}^2$$

$$SQ_T = SQ_R + SQ_E$$

Tal como no modelo regressão linear simples, as estimativas para os parâmetros são centradas, ou seja,

$$\boldsymbol{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

e as variâncias são dadas pelos elementos da diagonal principal da matriz $[\boldsymbol{X}^T\boldsymbol{X}]^{-1}$.

$\boldsymbol{\Sigma_{\hat{\beta}}} = \hat{\sigma}^2[\boldsymbol{X}^T\boldsymbol{X}]^{-1}$ é $\begin{bmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) & \cdots & cov(\hat{\beta}_0, \hat{\beta}_k) \\ cov(\hat{\beta}_1, \hat{\beta}_0) & var(\hat{\beta}_1) & \cdots & cov(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{\beta}_k, \hat{\beta}_0) & cov(\hat{\beta}_k, \hat{\beta}_1) & \cdots & var(\hat{\beta}_k) \end{bmatrix}$ com $\hat{\sigma}^2 = \frac{SQ_E}{n-p}$

Assim, a distribuição amostral dos parâmetros é $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}} \sim N(\boldsymbol{\beta}_j, \sigma^2 c_{j+1,j+1}); j = 0, 1, \ldots, k.$

Teste de significância e intervalos de confiança para os parâmetros da regressão

Hipóteses a testar:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Como os parâmetros têm distribuição Normal com média e variância dadas acima, ou seja,

$$\hat{\beta}_j \sim N\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right)$$

Assim, a estatística de teste é

$t = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \sim t_{(n-p)}, j = 0, 1, \ldots k$, onde $S_{\hat{\beta}_j} = \hat{\sigma}\sqrt{c_{j+1,j+1}}$

Com base nesta estatística podemos construir intervalos de confiança e efetuar testes de hipóteses aos parâmetros individuais de forma análoga à efetuada no modelo simples. Assim,

Intervalos de confiança $(1 - \alpha)$ para β_j é dado por

$$\left[\hat{\beta}_j - t_{(1-\alpha/2, n-p)} S_{\hat{\beta}_j} ; \hat{\beta}_j + t_{(1-\alpha/2, n-p)} S_{\hat{\beta}_j}\right].$$

4.3. Significado e avaliação da qualidade da regressão

a) Significado da regressão

De forma perfeitamente análoga à efetuada no contexto da regressão linear simples podemos efetuar uma ANOVA cuja tabela correspondente á a seguinte:

Tabela 13 : Tabela ANOVA de modelo de regressão múltipla

Fonte de variação	SQ	$g.l.$	MQ	F_{obs}	$P-value$
Regressão (explicada)	SQ_R	k	$MQ_R = \frac{SQ_R}{k}$	$\frac{MQ_R}{MQ_E}$	$P(F > F_{obs})$
Erros (não explicada)	SQ_E	$n-p$	$MQ_E = \frac{SQ_E}{n-p}$		
Total	SQ_T	$n-1$			

No modelo de regressão linear múltipla a ANOVA dá resposta ao teste

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs. $H_1 : \beta_j \neq 0$ para algum j

Rejeita-se H_0 para os valores elevados da estatística de teste

$$F = \frac{MQ_R}{MQ_E} \sim F_{k,n-p}$$

No fundo estamos a testar o significado da regressão num todo, ou seja, estamos a testar se tem significado considerar uma relação linear entre uma variável y e um conjunto de regressores $x_1, \dots, x_k$.

b) Avaliação da qualidade do modelo

Tal como na modelo de regressão linear simples, o coeficiente de determinação é definido por

$$R^2 = \frac{SQ_R}{SQ_T}, 0 \leq R^2 \leq 1$$

Ao adicionarmos variáveis regressoras ao modelo estamos sempre a aumentar o valor de R^2 e nem sempre essas variáveis são estatisticamente significativas. Assim, tal como na regressão simples, define-se o coeficiente de determinação ajustado para corrigir o viés do coeficiente:

$$R_a^2 = 1 - \frac{\frac{SQ_E}{n-p}}{\frac{SQ_T}{n-1}}, 0 \leq R_a^2 \leq 1$$

4.4. Previsão

Nesta secção apresenta-se a forma de utilizar o modelo de regressão linear múltipla para prever uma resposta, sendo esta a generalização dos métodos já apresentados para a regressão linear simples. Neste caso, o número de regressores é $k \geq 2$.

a) Previsão em média

Tal como no modelo da regressão linear simples, pretende-se estimar o parâmetro

$$\theta = E(y_i | x_{i1} = c_1, x_{i2} = c_2, \dots, x_{ik} = c_k) = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k$$

e o estimador de θ é

$$\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k.$$

Fazendo $c = \begin{bmatrix} 1 & c_1 & \dots & c_k \end{bmatrix}$, logo o valor esperado e variância do estimador de θ são os seguintes:

$$E(\hat{\theta}|X, c) = E(\hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k|X, c) = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k = \theta$$

$$Var(\hat{\theta}|X, c) = Var(c\hat{\beta}|X, c) = c\, Cov(\hat{\beta}|X, c)c^T = \sigma^2 c(X^T X)^{-1} c^T$$

Portanto, **erro padrão da previão em média** é dado por

$$s_{\hat{\theta}} = \hat{\sigma}\sqrt{c(X^T X)^{-1} c^T}$$

e o respetivo I.C. de previsão para θ com confiança $(1 - \alpha)$ é dado por

$$\left[\hat{\theta} - t_{1-\alpha/2} s_{\hat{\theta}}\, ; \hat{\theta} + t_{1-\alpha/2, n-2} s_{\hat{\theta}}\right].$$

b) Previsão pontual

Considere-se de novo que $x_{i1} = c_1, x_{i2} = c_2, \dots, x_{ik} = c_k$, e seja

$$y_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k + \varepsilon_0$$

Considere-se o previsor mínimo quadrado de y_0,

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$$

e o erro de previsão,

$$\varepsilon = y_0 - \hat{y}_0.$$

Utilizando a variável aleatória ε, vão estudar-se as propriedades estatísticas do previsor. Como

$$E(\varepsilon|X, c) = E(y_0 - \hat{y}_0|X, c) = 0$$

a variância de ε, condicionada por X e c, é dada por

$$Var(\varepsilon|x, c) = \sigma^2\{1 + c(X^T X)^{-1} c^T\}$$

Assim, o **erro padrão da previsão** é dado por

$$s_\varepsilon = \hat{\sigma}\sqrt{1 + c(X^T X)^{-1} c^T}$$

E o respetivo I.C. de previsão para y_0 com confiança $(1 - \alpha)$ é dado por

$$\left[\hat{y}_0 - t_{1-\alpha/2} s_\varepsilon\, ; \hat{y}_0 + t_{1-\alpha/2, n-2} s_\varepsilon\right].$$

4.5. Validação dos pressupostos da regressão

Os pressupostos do modelo de regressão linear múltipla são basicamente os mesmos apresentados para o modelo de regressão linear simples, e ainda o diagnóstico de pontos influentes (na estimação dos parâmetros) e análise de multicolinearidade.

4.5.1. Diagnóstico de pontos influentes

Um ponto de influente é uma observação que pode influenciar a construção do modelo de regressão. Os valores extremos ou os "pontos alavanca" têm potencial para serem pontos influentes, mas temos de investigar para avaliar o quanto eles são influentes. No entanto, uma observação

pode ser considerada um valor extremo e não ser um ponto influente. Da mesma forma, podemos ter pontos que influenciam na análise de regressão, mas não são valores extremos ou pontos de alavanca.

Um ponto diz-se influente se sua exclusão do ajuste da regressão causa uma mudança substancial na análise de regressão, por exemplo, nos valores ajustados ou nas estimativas dos coeficientes do modelo. Por isso, técnicas foram desenvolvidas para identificar essas observações influentes. ("Pontos Influentes - Análise de Regressão | Portal Action," n.d.). Existem vários critérios para identificação de pontos influentes (Hothorn & Everitt, 2014), e o software R implementa critérios que consideram amostras de grande dimensão (n grande). Em particular, define-se

a) DFBETA, mede a influência da observação i sobre o coeficiente de x_j, e a observação é influente se

$$DFBETA_{j(i)}: \left| \frac{\hat{\beta}_i - \hat{\beta}_{j(i)}}{\sqrt{QME_{(i)}c_{jj}}} \right| > \frac{2}{\sqrt{n}}, j = 0, 1, \dots, k$$

b) DFFITS, mede a influência que a observação i tem sobre seu próprio valor ajustado. Neste caso, medimos a influência da exclusão da i –ésima observação no seu previsto ou ajustado, e consideramos a observação influente se

$$DFFITS_i: \left| \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{QME_{(i)}h_{ii}}} \right| > 2\sqrt{\frac{k+1}{n-p}}$$

c) Distância de Cook, D_i, mede influência da observação i sobre todos os n valores ajustados $\hat{y}_i$, e consideramos a observação influente se

$$D_i = \frac{e_i^2 h_{ii}}{(k+1)QME(1-h_{ii})^2} > 1$$

onde,

- $\hat{y}_{i(i)}$ é previsão de y_i removendo a observação i,
- $QME_{(i)} = \sigma_{(i)}^2$ é variância do erro quando removendo a observação i,
- h_{ii} é diagonal principal do matriz chapéu, $H = \boldsymbol{X}[\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T$,
- $\hat{\beta}_{j(i)}$ é o valor estimado para β_j quando excluímos a observação (x_i, y_i).

4.5.2. Colinearidade

No modelo de regressão linear que envolva $k \geq 2$ variáveis explicativas, que estão fortemente correlacionadas entre si, tem-se o problema de colinearidade. Quando à correlação entre os regressores é alta, a eficiência dos parâmetros estimados diminui e em consequência disso, a variância da estimativa aumenta.

O indício da existência da colinearidade é quando o coeficiente de determinação, R^2 é bastante alto com poucos regressores significativos segundo o teste t, o valor da estatística F são muito grandes ou altas correlações dois a dois entre os regressores.

As consequências de existência da colinearidade, torna que os coeficientes da regressão não sejam estimados com precisão (erro padrão alto) e apresentam baixos níveis de significância, mesmo que sejam conjuntamente significativos e com coeficientes de determinação, R^2 elevado. Assim, dificulta a avaliação da importância relativa das variáveis independentes ao explicar a variância na variável independente.

A colinearidade pode ser diagnosticada pela análise da matriz correlação bivariadas. Quando mais de duas variáveis forem colineares, a matriz de correlações já não pode ser usada. Neste caso pode recorrer-se ao fator de inflação da variância (VIF) e, a colinearidade existe quando $VIF \geq 5$. (Mello, 2014)

$$VIF_j = \frac{1}{1 - R_j^2}$$

onde R_j^2 é o R^2 da regressão de X_j sobre as outras variáveis explicativas.

Numa situação destas, um procedimento para correção da colinearidade é exclusão das variáveis colineares. Neste caso, eliminar uma das variáveis e reestimar o parâmetros do modelo, através seleção de variáveis no modelo.

4.6. Seleção de variáveis numa regressão múltipla

Se existir apenas duas variáveis regressoras ($k = 2$), a seleção de variáveis é feita pela seleção de todos os modelos possíveis. Neste caso, através de comparações do coeficiente de determinação múltipla dos modelos possíveis.

Se existir mais de duas variáveis regressoras ($k > 2$), a seleção de variáveis é feita pela seleção automática, neste caso, é método de seleção "Stepwise", que consiste na combinação dos métodos *Backward (inclusão passo atrás)* e *Forward* (inclusão passo a frente). Geralmente parte-se de um modelo completo e, em cada passo avalia-se a exclusão e a inclusão de variáveis. Um critério muito utilizado para a comparação de modelos é o *Akaike information Criterion* (AIC) dado por

$$AIC = -2\ln(L|\beta) + 2p$$

em que $\ln(L|\beta)$ é o log natural da função de verossimilhança do modelo e p é o número de parâmetros do modelo. Quanto menor o valor do AIC, melhor é o ajuste do modelo aos dados.

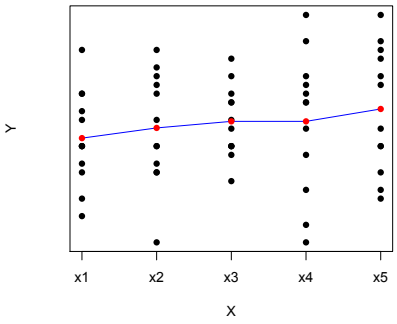
Neste trabalho, o método de seleção Stepwise é feito pelo recurso ao R.

5. Análise de variância (ANOVA)

A análise de variância (ANOVA) foi introduzida por R.A. Fisher (1918,1925,1935) e desenvolvida por H. Sheffé (1956). (Fonseca, 2001).

A análise de variância (ANOVA) é uma metodologia estatística com o objetivo de comparar três ou mais grupos, no que respeita à sua localização. Neste caso, é utilizada para verificar se existem diferenças significativas entre as médias dos grupos, que sejam resultado dos efeitos dos grupos.

Gráfico 14 : Gráfico de médias



Para aplicação da ANOVA, são necessárias algumas suposições, sendo eles:

1º - as amostras são aleatórias e independentes.

2º - as populações têm distribuição Normal (o teste é paramétrico).

3º - as variâncias populacionais são iguais (homogeneidade de variância entre grupos).

Quando os vários grupos são modelados pela distribuição Normal, com igual variância, utiliza-se uma técnica clássica paramétrica da análise de variância (ANOVA). Quando estes pressupostos não são validos podemos recorrer a uma técnica não-paramétrica, nomeadamente ao teste de Kruskal-Wallis. (Hall et al., 2011)

Quando se pretende fazer uma experiência na qual as observações se dividem em vários grupos classificados através de um só fator, diz-se que se tem uma "análise de variância com um fator" – *one-way ANOVA*.

Por outro lado, diz-se a análise de variância tem tantos "níveis" ou "efeitos" quantos grupos distintos se considerem. Se os grupos são determinados a partida, diz-se então que temos uma análise de variância com "efeitos fixos". E, se os grupos são retirados aleatoriamente de entre um conjunto alargado de possibilidades, diz-se então teremos uma análise de variância com "efeito aleatórios".

5.1. ANOVA com um fator e efeitos fixos

Num modelo de ANOVA com um fator e efeitos fixos (*ANOVA one-way and fixed effects*) têm-se k grupos e cada grupo tem n_j observações. No total temos $n = kn_j$ observações designadas por Y_{ij} onde $i = 1, \dots, n$ identifica a posição de cada observação dentro do grupo e $j = 1, \dots, k$ identifica o grupo.

O modelo de ANOVA com um fator e efeitos fixos pressupõe que cada observação pode ser modelada pela expressão:

$$Y_{ij} = \mu_j + \varepsilon_{ij} = \mu + \alpha_j + \varepsilon_{ij}; i = 1, \dots, n_j; j = 1, \dots, k$$

onde, Y_{ij} ⟶ observação i do grupo j;

μ_j ⟶ média do grupo j;

α_j ⟶ efeito (não aleatório) do grupo j;

ε_{ij} ⟶ erro aleatório da observação do i do grupo j.

Se queremos testar a hipótese de igualdade de localização (médias) dos vários grupos então queremos testar as hipóteses

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k = \mu \text{ vs. } H_1: \mu_j \neq \mu, \text{ para algum } j$$

A ideia de base para testar estas hipóteses é a seguinte: estima-se variância σ^2 por dois métodos diferentes, um que não depende da veracidade de H_0 e outro que depende da veracidade de H_0. Se H_0 for verdadeira, então as duas estimativas devem ser próximas, caso contrário, devem diferir significativamente.

A ANOVA começa por decompor a variabilidade das respostas em duas componentes fundamentais, isto é,

$$\sum_{i=1}^{n_j}\sum_{j=1}^{k}\left(Y_{ij} - \overline{Y}\right)^2 = \sum_{j=1}^{k} n_j\left(\overline{Y}_j - \overline{Y}\right)^2 + \sum_{i=1}^{n_j}\sum_{j=1}^{k}\left(Y_{ij} - \overline{Y}_j\right)^2 \Leftrightarrow SQ_T = SQ_E + SQ_D$$

onde, SQ_T é variabilidade total das observações Y_{ij} em relação à média global $\overline{Y}$.

SQ_E é variabilidade das observações **entre** grupos – corresponde à soma ponderada das variações das médias de cada grupo, $\overline{Y}_j$, em torno da média global, $\overline{Y}$ (a ponderação é feita pelo número de observações de cada grupo, n_j).

SQ_D é variabilidade das observações **dentro** dos grupos – corresponde à soma das variações das observações Y_{ij} dentro de cada um dos diferentes grupos.

No contexto de ANOVA, SQ_T tem $n-1$ graus de liberdades (só há um parâmetro a estimar que é a media global μ). Existem k grupos pelo que SQ_E tem $k-1$ graus de liberdade (estimar a media global μ). Finalmente, dentro de cada grupo temos n_j observações, ou seja, $n_j - 1$ graus de liberdade (há que estimar a média do grupo), como há k grupos temos $k\left(n_j - 1\right) = n - k$ graus de liberdade para SQ_D. Assim temos a seguinte partição dos graus de liberdade.

$$n - 1 = (k - 1) + (n - k)$$

Definimos ainda:

$$MQ_E = \frac{SQ_E}{k-1} \text{ e } MQ_D = \frac{SQ_D}{n-k}$$

onde, MQ_E é média da soma dos quadrados entre grupos e MQ_D é média da soma dos quadrados dentro dos grupos.

Os valores de MQ_E e MQ_D são as duas estimativas de σ^2 anteriormente referidas (sendo MQ_E aquele que depende da veracidade de H_0). Assim, quando H_0 é verdadeira, estes valores devem ser próximos e, consequentemente, a razão $\frac{MQ_E}{MQ_D}$ terá um valor próximo de 1. Se H_0 não for verdadeiro, então o valor de MQ_E será significativamente superior ao de MQ_D. Assim, a hipótese H_0 é rejeitada para valores elevados de $\frac{MQ_E}{MQ_D}$.

Logo, é precisamente a estatística de teste usada para efetuar o teste de hipóteses, isto é

$$F = \frac{MQ_E}{MQ_D} \sim F_{k-1,n-k}$$

Tipicamente uma ANOVA de efeitos fixos é resumida numa tabela do seguinte tipo.

Tabela 14 : Tabela ANOVA de efeitos fixos

Fonte de variação	SQ	$g.l.$	MQ	F_{obs}	$P-value$
Entre grupos	SQ_E	$k-1$	$MQ_E = \frac{SQ_E}{k-1}$	$\frac{MQ_E}{MQ_D}$	$P(F > F_{obs})$
Dentro dos grupos	SQ_D	$n-k$	$MQ_D = \frac{SQ_D}{n-k}$		
Total	SQ_T	$n-1$			

No caso de rejeitar H_0 é desejável efetuar comparação de médias de grupos duas a duas por forma a detetar diferenças estatisticamente significativas, uma das possibilidades para efetuar a comparação todos os pares de médias é através de **comparações múltiplas.**

Existem muitos métodos alternativos para efetuar comparações múltiplas. Iremos apenas considerar dois, o de **Tukey** e o de **Sheffé**.

Estes métodos consistem na construção de intervalos de confiança para todos os pares de médias de tal forma que o conjunto de todos os intervalos de confiança tenha uma certa confiança, $1-\alpha$.

Hipóteses a testar em cada comparação:

$$\begin{matrix} H_0: \mu_i = \mu_j \\ H_1: \mu_i \neq \mu_j \end{matrix}; \quad i \neq j; \quad i,j = 1, \dots, k$$

Região de rejeição

- O método de **Tukey:** $\left|\bar{X}_i - \bar{X}_j\right| \geq \frac{Q_{\alpha;k;n-k}\sqrt{MQ_E}}{\sqrt{n_j}}$ onde $Q_{\alpha;k;n-k}$ é o quantil de probabilidade $(1-\alpha)$ para distribuição Studentized Range (tabelada) com $(k,\ n-k)$ graus de liberdade. Este método é adequado para amostras de igual dimensão. (Mello, 2014)
- O método de **Scheffé**: $\left|\bar{X}_i - \bar{X}_j\right| > \sqrt{MQ_E\left(\frac{1}{n_i}+\frac{1}{n_j}\right)(k-1)F_{1-\alpha,k-1,\,n-k}}$ onde $F_{1-\alpha,k-1,\,n-k}$ é o quantil de probabilidade $(1-\alpha)$ para distribuição F de Snedecor (tabelada) com $(k-1,\ n-k)$ graus de liberdade. É adequada para qualquer dimensão de amostra.

5.2. ANOVA com um fator e efeitos aleatórios

Por vezes os grupos se consideram numa ANOVA são escolhidos aleatoriamente entre um conjunto vasto de probabilidades em vez de serem pré-determinados (fixos). O aspeto formal do modelo é o mesmo da ANOVA com efeitos fixos, mas a interpretação é pouca diferente.

$$Y_{ij} = \mu_j + \varepsilon_{ij} = \mu + \alpha_j + \varepsilon_{ij}; i = 1, \dots, n_j; j = 1, \dots, k$$

onde neste caso, α_j e ε_{ij} são aleatórios. Assume-se que os erros ε_{ij} são normalmente com média 0 e variância σ^2, $N(0, \sigma^2)$, os efeitos α_j também são normalmente com média 0 e variância σ^2, $N(0, \sigma_\alpha^2)$.

Num modelo de efeitos aleatórios, a forma mais apropriada de testar a igualdade das médias dos vários grupos é através das hipóteses

$$H_0: \sigma_A^2 = 0 \text{ vs. } H_1: \sigma_A^2 > 0$$

onde σ_A^2 é a variância do fator e σ^2 é a variância do erro.

A hipótese alternativa só pode ser $\sigma_A^2 > 0$, uma vez que a variância não pode ser negativa.

Note-se que, se os efeitos tiverem variância nula, então a média dos grupos não poderá variar.

Tal como no modelo de efeitos fixos, se H_0 é verdadeira, os valores de MQ_E e MQ_D são estimativas de σ^2. Assim, continuamos a usar a razão $\frac{MQ_E}{MQ_D}$ para testar as hipóteses.

Quando H_0 é rejeitada, faz sentido estimar a variância do fator, σ_A^2. A estimativas correspondente é dada por (Fonseca, 2001)

- Para grupos com a mesma dimensão ($n_1 = n_2 = \cdots = n_k$), $\sigma_A^2 = \frac{MQ_E - MQ_D}{n_j}$
- Caso os grupos tiverem dimensões diferentes, n_j deve ser substituir por

$$r = \frac{1}{k-1}\left[\sum_{j=1}^{k} n_j - \frac{\sum_{j=1}^{k} n_j^2}{\sum_{j=1}^{k} n_j}\right]$$

Neste modelo não analisamos as comparações múltiplas, devido à natureza aleatória dos grupos.

5.3. Validação dos pressupostos da ANOVA

Previamente à realização da ANOVA há que verificar dois pressupostos fundamentais: normalidade dos grupos (dentro de cada grupo as observações seguem uma distribuição Normal) e homogeneidade das variâncias dos grupos (a dispersão dos dados em torno de média é igual para todos os grupos).

Para verificar a normalidade podemos recorrer aos testes como já referido no modelo de regressão linear. Quanto à homogeneidade de variâncias aplica-se o "**teste de Bartlett**" e o "**teste de Levene**".

Validados os pressupostos, procedemos então à construção da tabela ANOVA, caso contrário proceda um teste alternativa da ANOVA não paramétrica.

Teste de Bartlett

É um teste paramétrico para comparação de duas ou mais variâncias populacionais. Suponha-se que a partir de k populações se extraíram k amostras aleatórias simples e independentes, com dimensões $n_j, j = 1, \dots, k$. Sejam $\sigma_j^2, j = 1, \dots, k$ as variâncias das k populações. Este teste é adequado quando se assume normalidade dos dados.

Hipóteses a testar são:

H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ (variâncias homogéneas)

H_1: $\sigma_i^2 \neq \sigma_j^2; i \neq j; i, j = 1, \dots, k$; para algum j

Estatística do teste (Mello, 2014)

$$B = \frac{2{,}3026\left[\sum_{j=1}^{k}(n_j - 1).\, ln\left(\frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{\sum_{j=1}^{k}(n_j - 1)}\right) - \sum_{j=1}^{k}(n_j - 1).\, lns_j^2\right]}{1 + \frac{1}{3(k-1)}.\left[\sum_{j=1}^{k}\frac{1}{n_j - 1} - \frac{1}{\sum_{j=1}^{k}(n_j - 1)}\right]} = \frac{C}{A}$$

onde, n_j é dimensão do grupo j e s_j^2 é variância do grupo j.

A estatística do teste, B, distribui-se segundo χ_{k-1}^2. Assim, para um nível e significância α, rejeita-se H_0 quando $B > \chi_{(1-\alpha);(k-1)}^2$. Quando $C < \chi_{(1-\alpha);(k-1)}^2$ não é necessário calcular A, podendo logo concluir-se que não se rejeita H_0 em favor de H_1.

Teste de Levene

É um teste paramétrico para testar a homogeneidade das variâncias.

Hipóteses a testar

H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

H_1: $\sigma_i^2 \neq \sigma_j^2; i \neq j; i, j = 1, \dots, k$; para algum j

Estatística do teste é dada por

$$W = \frac{n-k}{k-1} \times \frac{\sum_{j=1}^{k} n_j\left(\bar{Z}_j - \bar{Z}\right)^2}{\sum_{j=1}^{k}\sum_{i=}^{n_j}\left(Z_{ij} - \bar{Z}_j\right)^2} \sim F_{k-1,n-k}$$

Onde, $Z_{ij} = \left|x_{ij} - \bar{X}_j\right|; i = 1, \dots, n_j; j = 1, \dots, k;$

x_{ij} é observação i do grupo j;

$\bar{X}_j$ é média do grupo j;

$\bar{Z}_j$ é média dos valores de Z para o grupo j;

$\bar{Z}$ é média global dos valores de Z.

Se a variável não tem distribuição normal, então Z deve calcular-se por $Z_{ij} = \left|x_{ij} - \tilde{X}_j\right|$ onde $\tilde{X}_j$ é a mediana do grupo j.

Assim, rejeita-se H_0 quando $W \geq F_{1-\alpha;\, k-1,n-k}$. Caso contrário não se rejeita H_0.

5.4. ANOVA não paramétrica

É a alternativa da one-way ANOVA, quando os pressupostos de normalidade e homogeneidade são violados. Neste caso, é aplicar o teste Kruskal-Wallis.

Este teste é uma generalização, para $k > 2$ ammostras, do teste de Mann-Whitney que permite encontrar diferenças significativas entre os valores centrais (mediana) de 3 ou mais amostras independentes. Aplica-se a variáveis de nível pelo menos ordinal, sendo baseado em ordenações.

Antes de calcular a estatística de teste, devem-se ordenar todas as observações por ordem crescente e deve-se atribuir a cada uma a sua ordem na amostra global. No caso de empates (isto é, existirem valores repetidos) atribui-se a média aritmética das ordens que as observações teriam.

Hipóteses a testar

$H_0: \theta_1 = \theta_2 = \cdots = \theta_k$

$H_1: \theta_i \neq \theta_j; i \neq j; i, j = 1, \ldots, k$, para algum j

Estatística de teste (Reis et al., 2016)

$T = \frac{1}{S^2}\left(\sum_{i=1}^{k} \frac{{R_i}^2}{n_i} - \frac{n(n+1)^2}{4}\right)$ (com valores repetidos nas amostras)

esta estatística reduz-se a

$T^* = \frac{12}{n(n+1)}\sum_{i=1}^{k} \frac{{R_i}^2}{n_i} - 3(n+1)$ (sem valores repetidos nas amostras (não existe empates))

onde, $S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} R\left(X_{ij}\right)^2 - \frac{n(n+1)^2}{4}\right)$

$$n = \sum_{i=1}^{k} n_i$$

$R_i = \sum_{j=1}^{n_i} R\left(X_{ij}\right)$ é soma das ordens do grupo j;

$R\left(X_{ij}\right)$ o posto atribuído a X_{ij};

Os valores críticos para a rejeição da H_0 para $k = 3; n_i \leq 5,\ i = 1, 2, \ldots, k$ e não existe empates entre grupos, encontram-se na tabela "quantis da estatística de Kruskal-Wallis para pequenas amostras" em anexo (anexo 2.11). Caso contrário, a estatística de teste segue aproximadamente a distribuição qui-quadrado com $k - 1$ graus de liberdade. Ou seja, rejeitar-se H_0 se $T_{obs} > tabelados$ ou se $T_{obs} > \chi^2_{1-\alpha;\, k-1}$. E assim, tem se a região de rejeição é $[\ valor\ tabelado; +\infty\ [$ ou $[\ \chi^2_{1-\alpha;\, k-1}; +\infty\ [$.

No caso de rejeitar H_0 é desejável efetuar um procedimento de comparação múltipla por forma a detetar diferenças estatisticamente significativas entre si.

Hipótese a testar

$H_0: \theta_i = \theta_j$

$H_1: \theta_i \neq \theta_j; i \neq j; i, j = 1, \ldots, k$, para algum j

Região de rejeição

$$|\bar{R}_i - \bar{R}_j| \geq t_{(n-k;1-1/\alpha)} \sqrt{S^2 \frac{n-1-T}{n-k} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

onde $\bar{R}_i$ e $\bar{R}_j$ são respetivamente a média do grupo i e j.

Parte III Material de Apoio de EAD (slides, folhas práticas e atividades para R)

Introdução

Esta parte da tese apresenta os materiais de apoio para uso na sala de aula, incluindo os slides de apoio associados aos capítulos apresentados na parte II. Em todo o conteúdo, vão surgindo exemplos resolvidos (e respetiva explicação) à medida que os diversos assuntos vão sendo apresentados. Cada capítulo termina com exercícios, folhas práticas e atividades para aprendizagem/treino com software R.

A abordagem computacional da matéria é uma característica importante em análise de dados estatísticos. O computador, além de tornar a aprendizagem mais aliciante, pode funcionar como um laboratório para simular a experimentação, possibilitando aos estudantes adquirir maior intuição e compreender mais facilmente a matéria. Por outro lado, o computador permite resolver problemas cuja solução não pode ser obtida de forma analítica. Todo processo de aprendizagem torna-se mais efetivos quando a teoria é combinada com a prática.

Neste trabalho é aplicando o software R em resolução dos exemplos e exercícios propostos, uma vez que o R é um software livre, então pode ser obter gratuitamente. Para começarmos a trabalhar com o R é necessário baixá-lo na página do R Project da internet. Então, digite http://www.r-project.org na barra de endereços do seu navegador. Em seguida clique no link **download R** embaixo da página, que o levará à página do CRAN (Comprehensive R Archive Network). Escolha o sistema operativo do seu computador e clique em **base**.

Estatística e Análise de Dados
Slide de apoio 2019

Alexandrina Maria da Silva
alexandrinasilva@ua.pt

Faculdade de Ciências Exatas (FCE)
Universidade Nacional de Timor Lorosa'e (UNTL)

1

Conteúdos

I. Análise exploratória de dados
1. Revisão de conceitos de estatística descritiva
2. Organização de dados
3. Medidas amostrais

II. Análise de tabelas de contingência
1. Revisão de conceito de probabilidade
2. Teste Hipóteses
3. Tabelas contingência r x c
4. Tabelas contingência r x c x l

III. Regressão linear e análise de variância
1. Introdução
2. Correlação e regressão
3. Regressão linear simples
4. Regressão linear múltipla
5. Análise de variância

2

Distribuição de horário

DISTRIBUIÇÃO DE HORÁRIO				
Tópicos	**Subtópicos**	**Horas**	**Total horas**	**Nº. aula**
Análise exploratório de dados	1. Revisão de conceitos de estatística descritiva	2	10	5 aulas
	2. Organização de dados	2		
	3. Medidas amostrais	6		
Análise de tabelas de contingência	1. Revisão sobre probabilidades	4	38	19 aulas
	2. Teste de hipóteses	14		
	3. Tabelas contingência r x c	6		
	4. Tabelas contingência r x c x l	14		
Regressão linear e análise de variância (ANOVA)	1. Introdução	2	38	19 aulas
	2. Correlação e regressão			
	3. Regressão linear simples	14		
	4. Regressão linear múltipla	12		
	5. Análise de variância (ANOVA)	10		
Avaliação contínua	1º teste	2	4	2 aulas
	2º teste	2		
Total			**90**	**45 aulas**

Capitulo 1 Análise exploratória de dados

3

I. Análise exploratória de dados

1. Revisão de conceitos de estatística descritiva

O que é a estatística ?

A Estatística é um conjunto de técnicas que permite, de forma sistemática, para recolher, organizar, apresentar, resumir e interpretar os conjuntos de dados, com objetivo de tirar conclusões sobre a informação contida nesses dados.

Tipos fundamentais da estatística

- Estatística descritiva
- Inferência estatística

4

Estatística descritiva, é a etapa inicial da análise utilizada para descrever e resumir os dados, ou seja, consiste na recolha, apresentar, análise e interpretar de dados numéricos através da criação de instrumentos adequados: quadros, gráficos e indicadores numéricos.

Inferência estatística integra um conjunto de técnicas que permitem fazer ilações acerca de uma característica desconhecida da população. Especificamente, permitem estimar os valores característicos das populações de interesse e efetuar teste que validem, ou não, uma hipótese formulada sobre esses valores ou sobre a forma da distribuição da variável.

5

População ou universo

(Conjunto de todos os elementos relativos a um determinado fenômeno que possuem pelo menos uma caraterística em comum, a população pode ser finita ou infinita).

Amostra

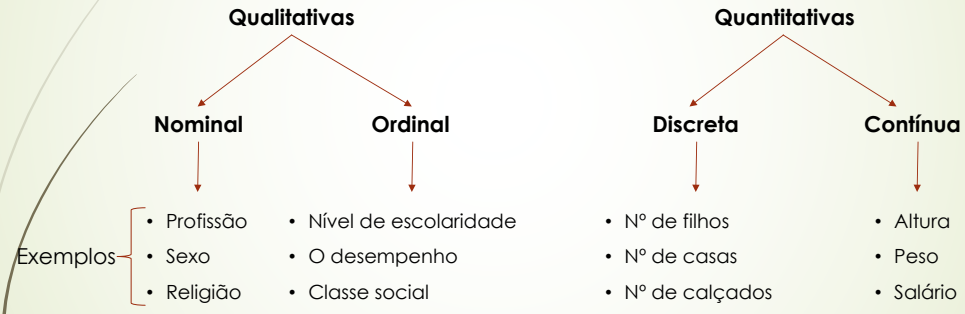
(É um subconjunto finito da população)

6

Variável

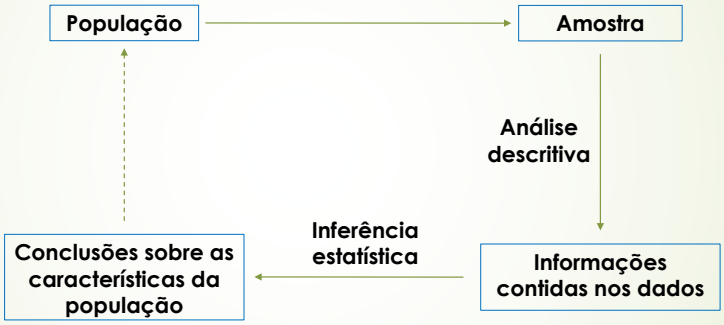
(É uma característica que pode tomar valores possíveis)

Tipos de variáveis



7

Etapas da Análise Estatística



8

Exercício 1:

Classifique os seguintes conjuntos de dados:

a) Conjuntos dos tempos esperados em uma fila
b) A quantidade de precipitação diária
c) Conjunto dos nome dos dia numa semana
d) Grau de satisfação
e) Conjuntos dos tamanho dos sapatos vendidos numa loja
f) Nível de educacional

9

2. Organização de dados

Uma maneira de organização de dados de forma estatisticamente seria através de:

- ❑ Tabela de distribuição de frequência
- ❑ Gráfico

Um procedimento que surge naturalmente antes de organizar os dados de forma estatisticamente é o da ordenação dos dados. Esta operação é baseada no *rank* das observações, e permite ordenar os dados quer por ordem crescente ou decrescente.

10

2.1. Tabela de frequências

Sejam $x_1^* < x_2^* < \cdots < x_k^*$ de k observações distintas (variável discreta) de ordem crescente numa amostra de dimensão n. Geralmente define-se os seguintes tipos de frequências:

- **Frequência absoluta:** $f_i \equiv$ número de vezes que se observou o valor x_i^* na amostra
- **Frequência relativa:** $f_{ri} = \frac{f_i}{n} \equiv$ proporção de valores iguais a x_i^* na amostra;
- **Frequência absoluta acumulada:** $F_i = f_1 + f_2 + \cdots + f_i = \sum_{j=1}^{i} f_j$
- **Frequência relativa acumulada:** $F_{ri} = f_{r1} + f_{r2} + \cdots + f_{ri} = \sum_{j=1}^{i} f_{rj} = \frac{1}{n}\sum_{j=1}^{i} f_j$

11

A **tabela de frequência** não é mais do que um quadro que concentra pelo menos um dos tipos de frequências da variável x_i numa amostra ou coleção de dados de dimensão n.

Distribuições de frequências de variáveis discretas

x_i^*	f_i	F_i	f_{ri}	F_{ri}
x_1^*	f_1	F_1	f_{r1}	F_{r1}
x_2^*	f_2	F_2	f_{r2}	F_{r2}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_k^*	f_i	n	f_{ri}	1
	$\sum_{i=1}^{k} f_i = n$		$\sum_{i=1}^{k} f_{ri} = 1$	

12

No caso de variáveis contínuas, é necessário definir classes para construir a distribuição de frequências. Para definir estas classes é necessário fixar o número de classes, a amplitude de classes e os limites dos intervalos. Logo, para proceder a contagem destes valores, seguimos os seguintes passos:

Passo 1º: Determinar a quantidade de classes (k)

- $k = 5$ para $n < 25$
- $k \approx \sqrt{n}$ para $n \geq 25$
- Formula de Sturges: $k = 1 + 3{,}3\log(n)$

Geralmente, na determinação o número de classe, a **regra de Sturges** fornece bons resultados e que, por isso, é uma boa escolha para iniciar a definição do número de classes a considerar.

13

Passo 2º: calcule a amplitude das classes (h)

- Amplitude do conjunto de dados : $R = X_{max} - X_{min}$
- Amplitude (largura) da classe: $h = R/k$; arredonde convenientemente

Passo 3º: Calcule os Limites das Classes

- 1ª classe: X_{min} até $X_{min} + h$
- 2ª classe: $X_{min} + h$ até $X_{min} + 2h$
- .....
- k ª classe: $X_{min} + (k-1)h$ até $X_{min} + kh$

Passo 4º: Ponto médio das classes

$$X_m = \frac{L_{inf} + L_{sup}}{2}$$

Passo 5º: construir tabela de frequências

X	f_i	F_i	f_{ri}	F_{ri}
$]L_0 , L_1]$	f_1	F_1	f_{r1}	F_{r1}
$]L_1 , L_2]$	f_2	F_2	f_{r2}	F_{r2}
⋮	⋮	⋮	⋮	⋮
$]L_{k-1} , L_k]$	f_k	n	f_{rk}	1
	$\sum_{i=1}^{k} f_i = n$		$\sum_{i=1}^{k} f_{ri} = 1$	

14

Exemplo 1

Os seguintes dados referem-se ao tempo gasto (em minutos) por 42 trabalhadores entre a sua casa e a local de trabalho no capital de Dili.

5	21	26	42	24	29	37
12	31	5	50	18	33	14
23	22	17	32	7	17	42
15	38	20	11	26	25	29
27	8	24	12	39	25	28
48	47	19	22	28	9	18

1. Construa um quadro de distribuição de frequências para os dados acima.
2. Construa também um quadro de distribuição de frequências depois de definir a amplitude das classes do modo que achar mais conveniente.

5	21	26	42	24	29	37
12	31	5	50	18	33	14
23	22	17	32	7	17	42
15	38	20	11	26	25	29
27	8	24	12	39	25	28
48	47	19	22	28	9	18

15

Soluções:

1. Ordenar os dados:

   5, 5, 7, 8, 9, 11, 12, 12, 14, 15, 17, 17, 18, 18, 19, 20, 21, 22, 22, 23, 24, 24, 25, 25, 26, 26, 27, 28, 28, 29, 29, 31, 32, 33, 37, 38, 39, 42, 42, 47, 48, 50

   Tabela de frequências:

x_i	f_i	F_i	f_{ri}	F_{ri}
5	2	2	0,05	0,05
7	1	3	0,02	0,07
8	1	4	0,02	0,09
9	1	5	0,02	0,11
11	1	6	0,02	0,13
12	2	8	0,05	0,18
14	1	9	0,02	0,20
15	1	10	0,02	0,22
17	2	12	0,05	0,27
18	2	14	0,05	0,32

Continuação ...

x_i	f_i	F_i	f_{ri}	F_{ri}
19	1	15	0,02	0,34
20	1	16	0,02	0,36
21	1	17	0,02	0,38
22	2	19	0,05	0,43
23	1	20	0,02	0,45
24	2	22	0,05	0,50
25	2	24	0,05	0,55
26	2	26	0,05	0,60
27	1	27	0,02	0,62
28	2	29	0,05	0,67
29	2	31	0,05	0,72

x_i	f_i	F_i	f_{ri}	F_{ri}
31	1	32	0,02	0,74
32	1	33	0,02	0,76
33	1	34	0,02	0,78
37	1	35	0,02	0,80
38	1	36	0,02	0,82
39	1	37	0,02	0,84
42	2	39	0,05	0,89
47	1	40	0,02	0,91
48	1	41	0,02	0,93
50	1	41	0,02	0,95
total	**42**		**1**	

5	21	26	42	24	29	37
12	31	5	50	18	33	14
23	22	17	32	7	17	42
15	38	20	11	26	25	29
27	8	24	12	39	25	28
48	47	19	22	28	9	18

16

Soluções:

2. Os passos de classificação dos dados:

   Passo 1º: Determinar a quantidade de classes (k), usando a formula de Sturges

$$k = 1 + 3{,}3\log(n) = 1 + 3{,}3\log(42) = 1 + 3{,}3 \times 1{,}62 = 6{,}36 \approx 7$$

   Passo 2º: Calcule a amplitude das classes (h)

$$R = X_{max} - X_{min} = 50 - 5 = 45$$

$$\Longrightarrow h = \frac{R}{k} = \frac{45}{7} = 6{,}42 \approx 7$$

   Passo 3º: Calcule os Limites das Classes

   - 1ª classe: 5 até 12
   - 2ª classe: 12 até 19
   - 3ª classe: 19 até 26
   - 4ª classe: 26 até 33
   - 5ª classe: 33 até 40
   - 6ª classe: 40 até 47
   - 7ª classe: 47 até 54

17

Passo 4º: Ponto médio das classes

$$X_{mi} = \frac{L_{inf} + L_{sup}}{2}$$

$$\Leftrightarrow X_{m1} = \frac{5 + 12}{2} = 8{,}5$$

$$\vdots$$

$$X_{m7} = \frac{47 + 54}{2} = 50{,}5$$

Passo 5º: construir tabela de frequências

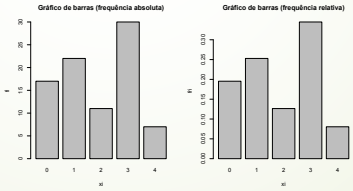
x_i	x_{mi}	f_i	F_i	f_r	F_r
[5, 12[	8,5	6	6	0,14	0,14
[12, 19[	15,5	8	14	0,19	0,33
[19,26 [	22,5	10	24	0,24	0,57
[26,33[	29,5	9	33	0,21	0,78
[33,40[	36,6	4	37	0,10	0,88
[40,47[	43,5	2	39	0,05	0,93
[47,54[	50,5	3	**42**	0,07	**1**
Total		**42**		**1**	

18

2.2. Representação gráfica da distribuição de frequência

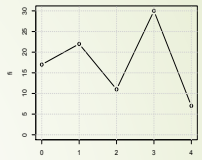
Neste caso, para as **variáveis qualitativas nominais**, são feitas através de **gráficos de barras** (muitas categorias) e **gráfico de setores** (poucas categorias). **Variáveis qualitativas ordinais** é através de **gráfico de barras e gráfico de linhas**. **Variáveis quantitativas discretas** é através de **gráfico de barras e gráfico de linhas**. E, **variáveis quantitativas contínuas** é através de **gráfico setores, histograma e polígono de frequências**.

a. **Gráfico de barras** (ou de colunas) é utilizado, em geral, para representar dados de uma tabela de frequências associadas a uma variável qualitativa. Nesse tipo de gráfico, cada barra retangular representa a frequência absoluta ou a frequência relativa da respetiva variável.

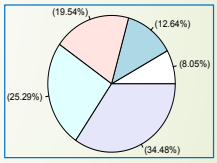


19

b. Gráfico de linhas (ou de segmentos) é utilizado, em geral, para representar a evolução dos valores de uma variável no decorrer do tempo.



c. Gráfico de setores, também conhecido como "**gráfico de pizza**", ou "**diagrama circular**" é utilizado, em geral, para representar partes de um todo. Para construir um gráfico de setores é necessário determinar o ângulo dos setores circulares correspondentes. Neste caso, a frequência total corresponderia aos 360º e a frequência ou a proporção de cada categoria corresponderia a um valor desconhecido em graus.



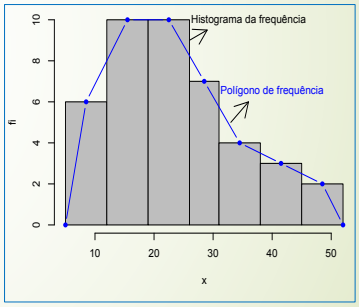
$$Graus\ de\ uma\ categoria = 360^0 \times \frac{freq.abs.da\ categoria}{dimensão\ da\ amostra} = 360^0 \times \frac{\boldsymbol{f_i}}{\boldsymbol{n}} = 360^0 \times f_{ri}$$

20

d. Histograma e polígono de frequências.

O **histograma** é uma representação constituída por uma sucessão de retângulos (barras) adjacentes em que cada um tem por base um intervalo de classe e a altura é igual à respetiva frequência (relativa ou absoluta) dessa classe.

O **polígono de frequência** é um gráfico de linhas onde são representadas as frequências (relativa ou absoluta) nos pontos médios das classes. Para fechar o polígono basta ligar a frequência associada ao ponto médio da classe extrema ao ponto de abcissa igual ao limite inferior (para a primeira classe) ou ao limite superior (para a última classe) e ordenada zero, tal como ilustrado na figura ao lado.

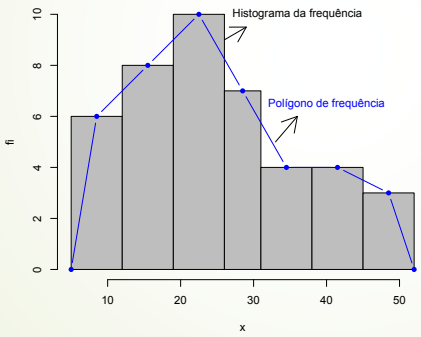


21

Exemplo 2

Representa graficamente os dados considerado no "***exemplo 1***" (*tempo é uma variável quantitativa contínuas*).

Soluções:



NOTA:

Neste caso, não representar no gráfico de setores, uma vez que existem muitos setores

22

3. Medidas amostrais

Medidas de localização

As medidas de localização indicam os valores da variável estatística onde os dados observados mais se concentram, estes designam-se por "***medidas de tendência central***".

As medidas de tendência central mais usadas são: **média aritmética**, **mediana**, **moda** e **quantis**.

Medidas de dispersão

As medidas de dispersão têm como objetivo descrever a variabilidade ou dispersão existente num determinado conjunto de dados.

As medidas de dispersão são: **amplitude amostral**, **amplitude interquartis**, **variância** e **desvio padrão**.

Medidas de assimetria (skewness)

Para além de da localização e dispersão tem por vezes interesse considerar a assimetria (ou enviesamento) dos dados.

23 **3.1. Medidas de localização**

a) **Média amostral** ou **média aritmética** é calculado por a soma de todos os valores observados dividida pelo número de observações. A média é definida pela expressão

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{k} f_i x_i^* = \sum_{i=1}^{k} f_{ri} x_i^* \ (variavel\ discreta)$$

$$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{k} f_i x_{mi} = \sum_{i=1}^{k} f_{ri} x_{mi} \quad (dados\ agrupados\ em\ classe \longrightarrow variável\ contínua)$$

onde:

k é o número de observações distintos (variável discreta) ou o número de classes.

f_i e f_{ri} são respetivamente a frequência absoluta e a frequência relativa da classe i

x_{mi} é o ponto médio da classe i

24 **b)** **Mediana**, é o valor da variável que ocupa a posição central de um conjunto de dados ordenados. A posição da mediana é dada pela expressão

$$Md = \begin{cases} x_{\frac{n+1}{2}:n} & se\ n\ for\ impar \\ \frac{x_{\frac{n}{2}:n} + x_{\frac{n}{2}+1:n}}{2} & se\ n\ for\ par \end{cases} \quad (variável\ discreta)$$

No caso variáveis contínua, a mediana é a classe cuja frequência aquela em que a frequência relativa acumulada atinge os 50%. O valor exato da mediana pode calcular-se através de expressão:

$$M_d = L_{inf} + \frac{\frac{n}{2} - F_{ant}}{f} \times h \quad (dados\ agrupados\ em\ classes)$$

onde: M_d = Mediana

L_{inf} = Limite inferior de classe mediana

F_{ant} = Frequência absoluta acumulada da classe anterior

f = Frequência absoluta da classe mediana

h = Amplitude da classe mediana

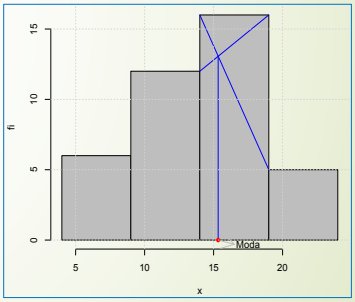
25

c) **Moda,** é o valor ocorre com maior frequência. A moda não tem de ser única pois pode haver mais do que um valor x_i^* com igual frequência sendo essa frequência máxima.

Exemplo, $x = 4,5,4,6,5,8,4,4,4 \implies M_o = 4$

No caso de variáveis contínua, a classe modal é a que tiver maior frequência absoluta. Pode determinar-se valor por aplicação de uma formula ou ilustrar por construção gráfica. A formula é dada por:

$$M_o = L_{inf} + \frac{\mathrm{d}_1}{\mathrm{d}_1 + \mathrm{d}_2} \times h$$



Onde,

M_o é Moda

L_{inf} é limite inferior de classe modal

d_1 é diferença da frequência absoluta da classe modal e da classe anterior

d_2 é diferença da frequência absoluta da classe modal e da classe posterior

h é amplitude da classe modal

26

d) **Mínimo** e **Máximo**

O mínimo é a observação com *rank* ascendente igual a 1, ou seja, $x_{1:n}$. O máximo é a observação com *rank* ascendente igual a n, correspondendo a $x_{n:n}$.

e) **Quantis**

Denomina-se por quantil de ordem p, $p \in (0,1)$, o valor real que Q_p que detém, à sua esquerda, (aproximadamente) $p\,x\,100\%$ das observações que compõem a amostra. Os mais utilizados são os **quartis**, os **decis** e os **percentis**.

- **Quartis,** são os quantis de ordens $p_i = \frac{i}{4}$, com $i = 1,2,3$, ou seja, $p = \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$. Os quartis dividem um conjunto de dados em 4 parte iguais. Assim, o número de quartis é 3, são respetivos 1º quartil (Q_1) ou quartil inferior, 2º quartil (Q_2) ou mediana e 3º quartil ou quartil superior (Q_3).
- **Decis,** são os quantis de ordens $p_i = \frac{i}{10}, i = 1, \dots, 9$, ou seja, $p = \{\frac{1}{10}, \frac{2}{10}, \dots, \frac{9}{10}\}$. Os decis dividem um conjunto de dados em 10 partes iguais.
- **Percentis,** são os quantis de ordens $p_i = \frac{i}{100}, i = 1, \dots, 99$, ou seja, $p = \{\frac{1}{100}, \frac{2}{100}, \dots, \frac{99}{100}\}$. Os percentis, dividem um conjunto de dados em 100 partes iguais.

27

Para facilitar a obtenção dos quantis, que se calculam a partir da amostra ordenada por ordem crescente. Assim, os quantis de ordem p, podemos calcular das seguintes formas:

$$Q_p = \begin{cases} x_{[np]+1:n} & se\ np\ não\ for\ inteiro \\ \frac{1}{2}\left(x_{np:n} + x_{np+1:n}\right) & se\ np\ for\ inteiro \end{cases} \qquad (variável\ discreta)$$

e

$$Q_p = L_{inf} + \left(\frac{np - F_{ant}}{f}\right) \times h \qquad (Variável\ contínua)$$

Onde $[np]$ representa a parte inteira de np.

Q_p é quantil de ordem p

L_{inf} é limite inferior de classe quantil de ordem p

n é total de observações

F_{ant} é frequência acumulada da classe anterior

h é amplitude da classe quantil de ordem p

f é frequência absoluta da classe quantil de ordem p

Para identificar a classe cuja frequência absoluta acumulada contem $n \times p$, isto é, contem o quantil em estudo.

28

3.2. Medidas de dispersão

a) **Amplitude amostral,** também chamada **amplitude total** ou **amplitude do intervalo de variação,** é diferença entre o máximo e o mínimo. Designa-se usualmente pela letra R devido a palavra inglesa "*Range*".

$$R = x_{max} - x_{min}$$

b) **Amplitude interquartil**, é a diferença entre o quartil inferior e o quartil superior. Ou seja, deve ser escrever na forma :

$$R_{IQ} = Q_3 - Q_1$$

29

c) Variância é a soma do quadrado das diferenças entre os valores da variável e a média, dividida pelo número total de observações. Nesta estatística, habitualmente denotada por s^2, quantifica a variabilidade dos dados em torno de uma característica central que é a média.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(x_i^* - \bar{x})^2 \quad (\text{variável discreta})$$

e

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(x_{mi} - \bar{x})^2 \quad (\textit{variável contínua})$$

onde, x_i é observação i

s^2 é variância amostral

$\bar{x}$ é média amostral (a estimar)

x_i^* observação i distinta

k é observações distintas (variável discreta) ou número de classe (variável contínuas)

x_{mi} é ponto médio da classe i

d) Desvio-padrão, é define-se como a raiz quadrada da variância, e se denota, naturalmente, por s.

30

3.3. Medidas de assimetria (skewness)

As medidas de assimetria indicam o grau de assimetria de uma distribuição. Há varias maneiras de analisar a simetria de uma distribuição amostral, algumas mais usadas são:

a) Coeficiente de assimetria

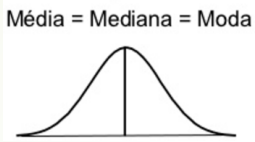
$$B = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} \quad (\textit{variável discreta}) \quad \text{e} \quad B = \frac{\frac{1}{n}\sum_{i=1}^{k} f_i(x_{mi} - \bar{x})^3}{s^3} \quad (\textit{variável contínua})$$

- Se $B = 0$, a distribuição é simétrica
- Se $B > 0$, há evidências de assimetria positiva
- Se $B < 0$, há evidências de assimetria negativa

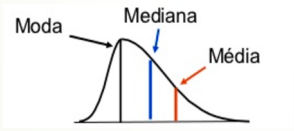
31

b) Comparando as três medidas de tendência central

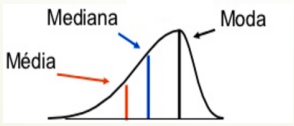
- $M_o = M_d = \bar{x} \Longrightarrow$ distribuição simétrica



- $M_o \leq M_d \leq \bar{x} \Longrightarrow$ distribuição assimétrica à esquerda ou assimetria positiva



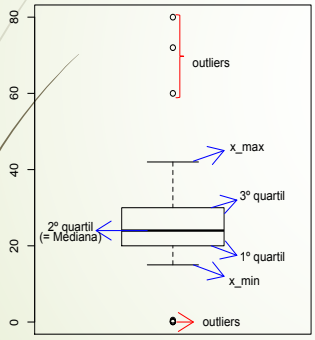
- $M_o \geq M_d \geq \bar{x} \Longrightarrow$ distribuição assimétrica à direita ou assimetria negativa



32

3.4. Representação gráfica de algumas medidas de localização e de dispersão

A maneira de apresentar graficamente e resumir algumas medidas de localização e dispersão é através de **Caixas de bigodes** (boxplot).



No *boxplot* é possível identificar as seguintes características:

- Localização: indicada pela mediana
- Dispersão: comprimento da caixa (AIQ) e do comprimento total entre os extremos dos bigodes (R)

Todas as observações que excedam os limites dos bigodes identificadas como *outliers*.

- *Outliers moderados,* são todas as observações que se situam para além de barreiras $Q_1 - 1{,}5 \times AIQ$ e $Q_3 + 1{,}5 \times AIQ$.
- se, além disso, ainda ultrapassarem as barreiras $Q_{0{,}25} - 3 \times AIQ$ e $Q_{0{,}75} + 3 \times AIQ$ os as observações designam-se *outliers severos ou extremos*

33 **Exemplo 3**

Considerando de cada um dos os dados no "**exemplo 1**", determinar:

a) Medidas de localização: Média, mediana, moda, mínimo e máximo, os quartis, 4º decil e 20º percentil

b) Medidas de dispersão: Amplitude de amostra (range), amplitude interquartis, variância e desvio-padrão

c) Faça uma caixa de bigode e mostrar algumas características sobres medidas de localização e dispersão.

d) O coeficiente e sinal de assimetria da distribuição

e) Compara os resultados da alínea a) e d), se são coincide.

34 **Soluções**

1. Em relação aos dados completo (dados reais):

a) Medidas de localização

➢ **Média**

$$\bar{x} = \frac{2\times5 + 1\times7 + \ldots + 2\times20}{42} = \frac{1025}{42} \approx 24{,}40$$

➢ **Mediana**

$n = 42 \longrightarrow par$

$$Md = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_{21} + x_{22}}{2} = 24$$

➢ **Moda:** neste caso, os dados têm multimodal

➢ **Mínimo e máximo**

$x_{min} = 5$ e $x_{max} = 50$

➢ **Quartis**

- $p = \frac{1}{4} \Rightarrow n\times p = \frac{42}{4} = 10{,}5 \Rightarrow \boldsymbol{Q_1} = \frac{x_{10}+x_{11}}{2} = \frac{15+17}{2} = 16$
- $p = \frac{1}{2} \Rightarrow n\times p = \frac{42}{2} = 21 \Rightarrow \boldsymbol{Q_2} = x_{21+1} = x_{22} = 24$
- $p = \frac{3}{4} \Rightarrow n\times p = \frac{126}{4} = 31{,}5 \Rightarrow \boldsymbol{Q_3} = \frac{x_{31}+x_{32}}{2} = \frac{29+31}{2} = 30$

➢ **4º decil**

$$p = \frac{2}{5} \Rightarrow n \times p = \frac{84}{5} = 16{,}8 \Rightarrow \boldsymbol{D_4} = \frac{x_{16}+17}{2} = \frac{20+21}{2} = 20{,}5$$

➢ **20º percentil**

$$p = \frac{1}{5} \Rightarrow n\times p = \frac{42}{5} = 8{,}4 \Rightarrow \boldsymbol{P_{20}} = \frac{x_8 + x_9}{2} = \frac{12+14}{2} = 13$$

35

b) Medidas de dispersão

- **Amplitude de amostra (range)**

  $R = x_{max} - x_{min} = 50 - 5 = 45$

- **Amplitude interquartil**

  $R_{IQ} = Q_3 - Q_1 = 30 - 16 = 14$

- **Variância**, sabendo que $\bar{x} = 15{,}47$

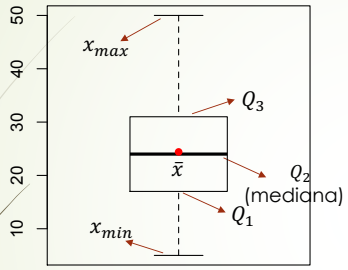
  $s^2 = \frac{1}{n-1}\sum_{i=1}^{6} f_i(x_i^* - \bar{x})^2 = \frac{1}{41}(2(5 - 24{,}40)^2 + \cdots + 1(50 - 24{,}40)^2) = \frac{1}{41} \times 5666{,}119$
  $= 138{,}198$

- **Desvio-padrão**

  $s = \sqrt{138{,}198} = 11{,}756$

36

c) Caixa de bigode (boxplot)



d) O coeficiente e sinal de assimetria

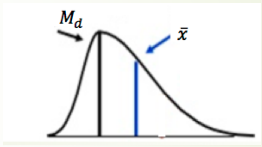
$$B = \frac{\frac{1}{n}\sum_{i=1}^{6} f_i(x_i - \bar{x})^3}{s^3} = \frac{\frac{1}{42} \times (567{,}522)}{1624{,}624} = 0{,}349$$

é assimétrica positiva

e) Compara os resultados da alínea a) e d)

Na alínea a): $\bar{x} = 24{,}40$; $M_d = 24$

Assim, $M_d < \bar{x}$



Logo, os resultados em a) e d) são concordantes com uma assimetria positiva da distribuição.

37

2. Em relação aos dados agrupados (classificados)

x_i	x_{mi}	f_i	F_i	f_r	F_r
[5, 12[	8,5	6	6	0,14	0,14
[12, 19[	15,5	8	14	0,19	0,33
[19,26 [	22,5	10	24	0,24	0,57
[26,33[	29,5	9	33	0,21	0,78
[33,40[	36,6	4	37	0,10	0,88
[40,47[	43,5	2	39	0,05	0,93
[47,54[	50,5	3	**42**	0,07	**1**
Total		**42**		**1**	

a) Medidas de localização

- **Média**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{7} f_i x_{mi} = \frac{1}{42}((6 \times 8{,}5) + \ \ldots + (3 \times 50{,}5) = \frac{1}{42} \times 1050 = 25$$

- **Mediana**

$n = 42 \Rightarrow \frac{n}{2} = \frac{42}{2} = 21 \longrightarrow$ classe mediana é 3ª classe

$$M_d = L_{inf} + \frac{\frac{n}{2} - F_{ant}}{f} \times h = 19 + \frac{21 - 14}{10} \times 7 = 23{,}9$$

- **Moda**

Classe modal é 3ª classe, porque tem maior frequência absoluta

$$M_o = L_{inf} + \frac{\mathrm{d}_1}{\mathrm{d}_1 + \mathrm{d}_2} \times h = 19 + \frac{10 - 8}{(10 - 8) + (10 - 9)} \times 7 = 23{,}67$$

38

x_i	x_{mi}	f_i	F_i	f_r	F_r
[5, 12[	8,5	6	6	0,14	0,14
[12, 19[	15,5	8	14	0,19	0,33
[19,26 [	22,5	10	24	0,24	0,57
[26,33[	29,5	9	33	0,21	0,78
[33,40[	36,6	4	37	0,10	0,88
[40,47[	43,5	2	39	0,05	0,93
[47,54[	50,5	3	**42**	0,07	**1**
Total		**42**		**1**	

- **Mínimo e máximo**

$x_{min} = 5$ e $x_{max} = 50$

- **Quartis**

- $p = \frac{1}{4} \Rightarrow n \times p = \frac{42}{4} = 10{,}5$, a classe 1° quantil é 2ª classe

$$\boldsymbol{Q_1} = Q_{\frac{1}{4}} = L_{inf} + \left(\frac{np - F_{ant}}{f}\right) \times h = 12 + \left(\frac{10{,}5 - 6}{8}\right) \times 7 = 15{,}94$$

- $p = \frac{1}{2} \Rightarrow n \times p = \frac{42}{2} = 21$, a classe de 2° quantil é 3ª classe

$$\boldsymbol{Q_2} = Q_{\frac{1}{2}} = L_{inf} + \left(\frac{np - F_{ant}}{f}\right) \times h = 19 + \left(\frac{21 - 14}{10}\right) \times 7 = 23{,}9 = M_d$$

- $p = \frac{3}{4} \Rightarrow n \times p = \frac{126}{4} = 31{,}5$, a classe de 3° quantil é 4ª classe

$$\boldsymbol{Q_3} = Q_{\frac{1}{2}} = L_{inf} + \left(\frac{np - F_{ant}}{f}\right) \times h = 26 + \left(\frac{31{,}5 - 24}{9}\right) \times 7 = 31{,}83$$

39

x_i	x_{mi}	f_i	F_i	f_r	F_r
[5, 12[	8,5	6	6	0,14	0,14
[12, 19[	15,5	8	14	0,19	0,33
[19,26 [	22,5	10	24	0,24	0,57
[26,33[	29,5	9	33	0,21	0,78
[33,40[	36,6	4	37	0,10	0,88
[40,47[	43,5	2	39	0,05	0,93
[47,54[	50,5	3	**42**	0,07	**1**
Total		**42**		**1**	

- **4º decil**

$$p = \frac{4}{10} = \frac{2}{5} \Rightarrow n \times p = \frac{84}{5} = 16{,}8$$

$$\boldsymbol{D_4} = Q_{\frac{2}{5}} = L_{inf} + \left(\frac{np - F_{ant}}{f}\right) \times h = 19 + \left(\frac{16{,}8 - 14}{10}\right) \times 7 = 20{,}96$$

- **20º percentil**

$$p = \frac{20}{100} = \frac{1}{5} \Rightarrow n \times p = \frac{42}{5} = 8{,}4$$

$$\boldsymbol{P_4} = Q_{\frac{1}{5}} = L_{inf} + \left(\frac{np - F_{ant}}{f}\right) \times h = 12 + \left(\frac{8{,}4 - 6}{8}\right) \times 7 = 14{,}1$$

40

x_i	x_{mi}	f_i	F_i	f_r	F_r
[5, 12[	8,5	6	6	0,14	0,14
[12, 19[	15,5	8	14	0,19	0,33
[19,26 [	22,5	10	24	0,24	0,57
[26,33[	29,5	9	33	0,21	0,78
[33,40[	36,6	4	37	0,10	0,88
[40,47[	43,5	2	39	0,05	0,93
[47,54[	50,5	3	**42**	0,07	**1**
Total		**42**		**1**	

b) Medidas de dispersão

- **Amplitude de amostra (range)**

$$R = x_{max} - x_{min} = 50 - 5 = 45$$

- **Amplitude interquartis**

$$R_{IQ} = Q_3 - Q_1 = 31{,}83 - 15{,}94 = 15{,}89$$

- **Variância**, sabendo que $\bar{x} = 25$

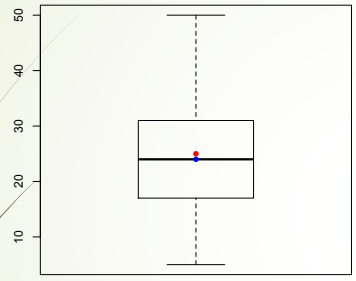
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{7} f_i(x_{mi} - \bar{x})^2 = \frac{1}{41}(6(8{,}5 - 25)^2 + \cdots + 3(50{,}5 - 25)^2) = \frac{1}{41} \times 5764{,}5 = 140{,}60$$

- **Desvio-padrão**

$$s = \sqrt{140{,}60} = 11{,}86$$

41

c) Caixa de bigode (boxplot)



d) O coeficiente e sinal de assimetria

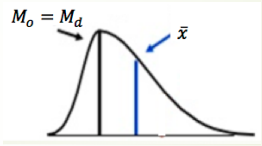
$$B = \frac{\frac{1}{n}\sum_{i=1}^{k} f_i(x_{mi} - \bar{x})^3}{s^3} = \frac{\frac{1}{42} \times (35343)}{1667{,}12} = 0{,}50$$

é assimétrica positiva

e) Compara os resultados da alínea a) e d)

Na alínea a): $\bar{x} = 25$, $M_d = M_o = 23{,}67$

Assim, $M_o = M_d < \bar{x}$



Logo, os resultados em a) e d) são concordantes com uma assimetria positiva da distribuição.

42

Exercício 2:

1. Considere os dados de idades dos 30 funcionários na empresa A seguintes:

26	33	22	27	34	26	22	32	22	27
27	22	27	26	20	32	36	20	36	27
20	32	20	27	36	22	26	32	27	36

a) Identifique a categoria dos dados.

b) Ordenar e construir a tabela de frequências dos dados distintos.

c) Representar graficamente (gráfico de barras para frequências absoluta e frequência relativa, gráfico de setores, gráfico de linha).

d) Calcular as medidas de localização: Média, mediana, moda, mínimo e máximo, os quartis, 5º decil e 25º percentil.

e) Calcular as medidas de dispersão: Amplitude de amostra (range), amplitude interquartis, variância e desvio-padrão.

f) Construir a caixa de bigode.

g) Determinar o coeficiente de assimetria e faça o comentário em relação a média, mediana, moda.

43

2. Considere os seguintes dados de 39 observações, de vendas mensais, em certas u.m., de uma empresa têxtil :

4,8	7,3	7,9	8,5	10,7	14,2	14,3	16,9	19,0	19,1
19,6	21,0	22,7	24,0	25,4	28,3	28,3	28,8	31,0	32,6
33,3	33,9	37,0	40,4	44,8	44,8	47,1	47,8	48,6	50,2
51,0	57,6	64,4	65,3	66,2	72,5	73,4	73,4	84,0	

a) Classificar os dados em classes e representar numa tabela das frequências.
b) Construir a histograma e polígono da frequência.
c) Calcular as medidas de localização: Média, mediana, moda, mínimo e máximo, os quartis, 5º decil e 25º percentil.
d) Calcular as medidas de dispersão: Amplitude de amostra (range), amplitude interquartis, variância e desvio-padrão.
e) Faça a representação de caixa de bigode.
f) Determinar o coeficiente e sinal de assimétrica e verifique se o resultado obtido são coincide em relação a comparação de média, mediana e moda.

44

Bibliografia

1. Mello, F.M., 2014. *Dicionário de Estatística. 673 entradas Índice remissivo em Português e inglês. Edições Sílabo, Lisboa*
2. Fonseca, J., 2001. *Estatística Matemática*. Vol. 2. Edições sílabo, Lisboa
3. Reis, E., 2009. *Estatística Descritiva*. 7ª Edição, Revista e Corrigida, Lisboa
4. Hall, A. Neves, C. e Pereira, A., 2011. *Grande Maratona de Estatística no SPSS*. Escolar Editora, Lisboa.
5. Magalhães, F. M. de, Oliveira, C. T. de, & Silva, E. Sá da. 2017. *estatística Descritiva Aplicada à Gestão – Uma Análise Exploratória dos Dados. Vida Económica. Porto.*

Folha prática 1

Universidade Nacional de Timor Lorosa'e (UNTL)

Faculdade de Ciências Exatas (FCE)

Estatística e Análise de Dados ano letivo 2019, 1º semestre

Análise Exploratória de Dados

1. Pesquisando a altura (em cm) de um grupo de pessoas obtivemos seguintes dados:

156	164	146	163	162
166	163	165	164	167
165	160	165	159	167
175	175	167	158	155
165	150	165	163	152

Em relação aos dados:

a) Identificar o tipo dos dados.

b) Construir a tabela de frequências dos dados acima e representa graficamente (gráfico de barras de frequência absoluta e relativa, gráfico de linhas, gráfico de setores e o boxplot).

c) Qual o número e a percentagem das pessoas que têm altura inferior a 159 cm?

d) Qual o número e a percentagem das pessoas que têm altura igual ou superior a 162 cm?

e) Organize a tabela de frequência agrupados em 6 classes, e construir o histograma e o polígono da frequência e gráfico de setores ou circular.

f) Determine (para dados não agrupados e dados agrupados em classes):

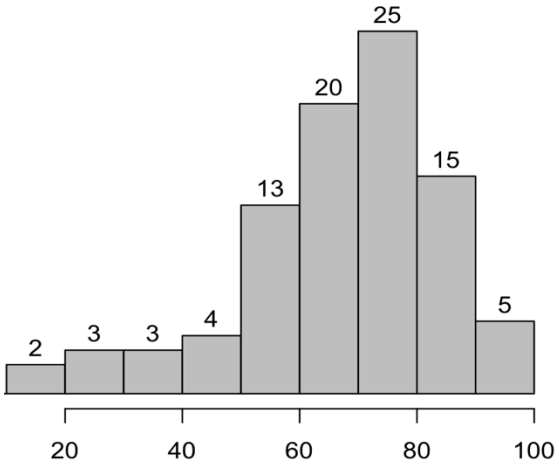
   i) Medidas tendência central: Média, mediana, moda e média aparada de 5%.

   ii) Medidas de localização relativa: mínimo e máximo, os quartis, 3º e 8º decil e 40º e 70º percentil.

   iii) Medidas de dispersão: Amplitude de amostra (range), amplitude interquartis, variância e desvio-padrão.

   iv) Coeficiente e sinal de assimétrica da distribuição.

2. Dado o histograma abaixo, calcular a mediana, moda e o 3º quartil.



3. Dois grupos de 20 estudantes foram selecionados ao caso para participar numa experiência que consiste em aprender o significado de palavra numa língua que não conhecem. Durante uma hora os estudantes tentarem aprender o maior número de palavras.
   No grupo I os estudantes trabalharam isoladamente. No grupo II os estudantes trabalharam aos pares procurando-se certificar mutuamente que iam aprendendo as palavras. Em seguida foi efetuando um teste para determinar o número de palavras aprendidas por cada aluno, tendo-se obtido os seguintes resultados:

Grupo I	24	14	16	17	18	23	14	15	15	17
	18	16	17	19	20	21	20	19	19	18
Grupo II	21	22	25	21	20	18	20	17	16	14
	17	15	18	23	17	19	15	23	19	20

   Estuda as características descritivas para os dados, representa graficamente e faça o comentário relativamente os resultados obtidos.

4. No âmbito de um estudo sobre o rendimento disponível mensal de duas populações, respetivamente as lojas de matérias de construção (A) e lojas de roupas (B) em Díli, foram extraídas amostras de 10 elementos de cada uma delas. Os resultados em milhares de dólares foram seguintes:

A	12	40	24	45	68	32	68	23	34	45
B	14	61	24	68	41	62	14	67	51	70

   a) Crie uma tabela de frequências relativamente a cada um dos casos, depois de classificar os dados utilizado os intervalos 10 – 30, 30 – 50 e 50 – 70.
   b) Considerar a média aritmética como indicador explicite qual a população com maior rendimento?

c) Represente graficamente a determinação da moda e da mediana e diga, exclusivamente pelo exame das figuras, se as conclusões tomadas a partir das mesmas corroboram a que pode tirar da alínea anterior.

d) Verifique se a conclusão na alínea b é corroborada pela análise comparada dos diagramas de extremos e quartis ("caixa de bigodes") destas distribuições.

5. Os salários pagos numa empresa seguradora (em u.m.) estão distribuídos segundo o seguinte quadro:

Classes de salários	**fi**
Menos de 25	25
25 – 40	75
40 – 60	200
60 – 120	150
120 – 250	50

a) Qual o salario médio na empresa?

b) Construa o histograma e o polígono de frequências.

c) Determine assimetria da distribuição usando a comparação das medidas de tendência central.

d) Determine a coeficiente de assimetria da distribuição.

e) Verifique que o seu sinal obtido em alínea c e alínea d estão em concordância com o histograma obtida.

f) Qual a percentagem de empregados da empresa que aufere salários superiores a 50 u.m.?

6. Considere os seguintes dados de chegada os estrangeiros ao Aeroporto Internacional de Díli, por país de origem, no ano de 2010 até 2013 são seguintes:

Países	2010	2011	2012	2013
Austrália	11 262	12 419	12 138	12 817
Brasil	803	978	1 722	707
China	2 659	3 464	4 972	4 346
EUA	1 720	2207	2 211	2 130
Filipinas	2 177	2 413	3 842	3 936
India	2 027	1 451	862	738
Indonésia	6 744	11 179	15 303	17 520
Japão	1 208	1 232	1 211	1 438
Malásia	1 756	1 829	1 944	1 455
Nova Zelândia	800	711	815	737
Paquistão	399	449	313	90
Portugal	996	5 916	6 130	5 894
Reino Unido	929	1 002	915	489
Singapura	1 495	1 519	1 381	1 457
Outros países	4 850	3 821	3 758	24 118
Total	**39 825**	**50 590**	**57 517**	**77 135**

Fonte/Source: Departamento de Imigração, Polícia Nacional de Timor-Leste
Immigration Department, Timor-Leste National Police

a) Calcule a média de chegada os estrangeiros de cada ano.
b) Qual o valor médio de chegada de cada país em destes 4 anos. Compare os valores médios e determine qual é o país que tem maior chegada em Díli neste período.
c) Determine a mediana e a moda e interpreta os resultados em relação às chegadas os estrangeiros em cada ano.
d) Qual a percentagem de chegada de Portugal e do Reino Unido dentro de total chegadas dos estrangeiros em 2012?
e) Construa o gráfico de barras respetivamente a chegada em cada ano.

Atividade para aprendizagem com R

```
                          ############################
                          #### VARIÁVEIS DISCRETAS ####
                          ############################

xi<-c(26,27,20,33,22,32,22,27,20,27,
       26,27,34,20,36,26,32,22,22,36,
       26,32,20,32,22,36,27,27,27,36)

n<-length(xi);n                  # dimensão de amostra
sort(xi)                         # ordenar os dados
min(xi);max(xi)
#range(xi)
summary(xi)
x_bar<-mean(xi);x_bar            # média amostral
mean(xi,trim=0.1)                # média aparada de 10 %
Md<-median(xi);Md                # mediana
quantile(xi,c(0.25,0.5,0.75))    # quartis
quantile(xi,c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9))  # decis
var(xi)                          # variância
#var<-sum((x-mean(x))^2)/(n-1);var
sd(xi)                           # desvio padrão
#sqrt(var)
round(sd(xi),2)                  # arredondar o valor até 2 casas decimais

## Representação numa tabela de frequência ##

fi<-table(xi)                    # tabela de frequência absoluta
Fi<-cumsum(fi)                   # frequência absoluta acumulada
#n<-sum(fi);n
fri<-fi/n                        # frequência relativa
Fri<-cumsum(fri)                 # frequência relativa acumulada
cbind(fi,Fi,fri,Fri)             # tabela de frequências
Mo<-names(fi[fi==max(fi)]);Mo    # moda

## Gráficos ##

barplot(fi)                      # gráfica de barra
pie(fi)                          # gráfica circular
plot(fi,type="b")                # gráfico de linha


# instala o package fBasics antes de aplicar o comando skewness #

skewness(xi)                     # coeficiente de assimetria
basicStats(xi)
```

```
                          ############################
                          #### VARIÁVEIS CONTÍNUAS ####
                          ############################


xi<-c(4.8,7.3,7.9,8.5,10.7,14.2,14.3,16.9,19.0,19.1,
       19.6,21.0,22.7,24.0,25.4,28.3,28.3,28.8,31.0,32.6,
       33.3,33.9,37.0,40.4,44.8,44.8,47.1,47.8,48.6,50.2,
       51.0,57.6,64.4,65.3,66.2,72.5,73.4,73.4,84.0)

## Agrupar os dados ##

k<-nclass.Sturges(xi);k                 # número de classe
#k<-1+3.3*log(42,10);k
R<-max(xi)-min(xi);R                    # amplitude total
h<-R/k;h                                # amplitude de classe

## Dividindo as observações em intervalos ##

intervalo <- cut(xi,breaks=c(4.8,16.8,28.8,40.8,52.8,64.8,76.8,88.8),right=FALSE)
intervalo

fi<-table(intervalo);fi                 # frequência absoluta
n<-sum(fi);n                            # dimensão da amostra
Fi<-cumsum(fi)                          # frequência absoluta acumulada
fri<-fi/n                               # frequência relativa
Fri<-cumsum(fri)                        # frequência relativa acumulada
cbind(fi,Fi,fri,Fri)                    # tabela de frequências

cl_Mo<-names(fi[fi==max(fi)]);cl_Mo     # classe modal

## histograma e polígono de frequência ##

par(mfrow=c(1,2))
h<-hist(xi,breaks=c(4.8,16.8,28.8,40.8,52.8,64.8,76.8,88.8),xlab="x",ylab="fi",
 freq=TRUE,right=FALSE,col="gray",main="")            # histograma
lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0),
 type = "b",col="blue",pch=16)                        # polígono de frequência

## ilustração gráfica de moda ##

h<-hist(xi,breaks=c(4.8,16.8,28.8,40.8,52.8,64.8,76.8,88.8),xlab="x",ylab="fi",
freq=TRUE,right=FALSE,col="gray",main="")
segments(16.8,7,28.8,10,col="red")
segments(28.8,7,16.8,10,col="red")
segments(22.8,0,22.8,8.5,col="red")
arrows(22.8,0,18,-0.3,col="green")
text(13.2,-0.25,expression(paste('Moda')),col="dark green")
```

Capitulo 2 Análise de tabelas de contingência

1

II. Análise de tabelas de contingência

Tabela de contingência é uma representação de dados, quer do tipo qualitativo ou quer do tipo quantitativo que podem ser classificados segundo dois critérios. Nesta tabela as linhas correspondem a um dos critérios e as colunas correspondem ao outro critério. No interior da tabela, as células correspondem ao número de observações O_{ij}, que satisfazem ambos os critérios. Uma tabela contingência, com $\boldsymbol{r}$ linhas e $\boldsymbol{c}$ colunas diz-se que tem dimensão $r \times c$, e designa-se por tabela de contingência *bidimensional,* com seguinte aspeto:

2

Tabela de contingência $\boldsymbol{r \times c}$

		$\boldsymbol{B}$	Totais
		$B_1 \ \ldots \ B_c$	
$\boldsymbol{A}$	A_1	$O_{11} \ldots O_{1c}$	$\boldsymbol{O_{1.}}$
	$\vdots$	$\vdots \quad \vdots$	$\vdots$
	A_r	$O_{r1} \ldots O_{rc}$	$\boldsymbol{O_{r.}}$
Totais		$\boldsymbol{O_{.1} \ldots O_{.c}}$	$\boldsymbol{n}$

Esta tabela de contingência bidimensional contém:

- uma amostra aleatória de dimensão $\boldsymbol{n}$, classificada relativamente a duas variáveis $\boldsymbol{A}$ e $\boldsymbol{B}$. (Quantitativa ou qualitativa),
- $\boldsymbol{r}$ linhas (categorias de variável $\boldsymbol{A}$);
- $\boldsymbol{c}$ colunas (categorias de variável $\boldsymbol{B}$);
- $\boldsymbol{rc}$ células (cruzamento das duas variáveis);
- $\boldsymbol{O_{r.}}$ (totais marginais das categorias de $\boldsymbol{A}$);
- $\boldsymbol{O_{.c}}$ (totais marginais das categorias de $\boldsymbol{B}$).

Algumas questões a responder:

- Será que A e B são independentes?
- Será que $A_1, \ldots, A_r$ tem a mesma distribuição de probabilidade para B?
- Será que a distribuição de probabilidade, por exemplo, de A_1 se ajusta a alguma distribuição conhecida?

3

A seguinte tabela de dupla entrada contem informação acerca de duas variáveis: sexo (linhas) e cor favorita (colunas) .

Sex \ Cor. fav.	Preta	Branca	Vermelha	Azul	Tot.
F	48	12	33	57	150
M	35	46	42	27	150
Total	83	58	75	84	300

Número de mulheres

Número de homens

Número total de pessoas (tamanho de amostra)

Número de pessoas que têm cor favorita preta

Número de pessoas que têm cor favorita branca

Cada célula corresponde ao número de pessoas com uma certa cor favorita e um certo sexo

Número de pessoas que têm cor favorita azul

Número de pessoas que têm cor favorita vermelha

4

As possibilidades de uma pessoa ter uma cor favorita determinada, podemos calcular através de **probabilidade**.

1. **Revisão de conceito de probabilidade**

Considere-se os seguintes elementos:

a) Espaço amostral ou espaço de amostral universal ou espaços dos resultados, é o conjunto de todos os resultados possíveis de ocorrência de um evento, simbolicamente representada pela: $\boldsymbol{S}$ ou $\boldsymbol{\Omega}$. O número de elementos do espaço amostral denotada por $\boldsymbol{n(S)}$ ou $\boldsymbol{n(\Omega)}$ e $\boldsymbol{n(\Omega) \neq 0}$.

b) Evento é qualquer subconjunto de um espaço amostral, denotada por qualquer uma letra maiúscula. O número de elementos de um evento é denotado por $\boldsymbol{n(.)}$.

5

c) Probabilidade de um acontecimento (evento) é quociente entre o número de casos favoráveis ao acontecimento e o número de casos possíveis, supondo que todos os casos são igualmente possíveis. Ou seja:

$$\boldsymbol{P}(.) = \frac{\sharp\ casos\ favoráveis}{\sharp\ casos\ possíveis} = \frac{\boldsymbol{n}(.)}{\boldsymbol{n}(\boldsymbol{\Omega})}, \qquad 0 \leq P(.) \leq 1$$

d) Classificação de eventos

- Evento certo, quando ele possui todos os elementos do espaço amostral. Ou seja, $\boldsymbol{n}(.) = \boldsymbol{n}(\boldsymbol{\Omega})$. Neste caso, $\boldsymbol{P}(.) = \boldsymbol{1}$
- Evento impossível, quando número de casos favoráveis é zero. ou seja $\boldsymbol{n}(.) = \boldsymbol{0} \Longrightarrow P(\emptyset) = 0$
- Evento complementar, dado um evento A num espaço amostral $\boldsymbol{\Omega}$. O complementar do evento $\bar{A}$ são todos os elementos do espaço amostral $\boldsymbol{\Omega}$ que não estão contidos em A, então temos que $\bar{A} = \boldsymbol{\Omega} - \boldsymbol{A}$ e ainda $\boldsymbol{\Omega} = \bar{A} + \boldsymbol{A}$.

  A probabilidade de evento complementar é:

$$P(\bar{A}) = P(\boldsymbol{\Omega} \backslash A) = P(\boldsymbol{\Omega}) - P(A) = 1 - P(A)$$

6

- Evento união. Dados dois eventos A e B de um espaço amostral $\boldsymbol{\Omega}$. O número de elementos de $A \cup B$ é igual à soma do número de elementos de A com o número elemento de elemento de B, menos uma vez o número de elementos de $A \cap B$ que foi contado duas vezes, assim temos:

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

A probabilidade de ocorrência A ou B é dada por:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Se A e B são dois eventos disjuntos ou mutuamente exclusiva. Ou seja,

$$A \cap B = \emptyset \Longrightarrow P(A \cap B) = 0$$

Logo, a sua probabilidade é simplificada por:

$$P(A \cup B) = P(A) + P(B)$$

7

- Evento intersecção. Dados dois eventos A e B de um espaço amostral $\boldsymbol{\Omega}$. O número de elementos de $A \cap B$ é igual o número de elementos simultâneos em A e B.

  A probabilidade de ocorrência simultânea de A e B, é dada por:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \Leftrightarrow \quad P(A \cap B) = P(B) \times P(A|B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \Leftrightarrow \quad P(A \cap B) = P(A) \times P(B|A)$$

Se A e B são eventos independentes, a probabilidade de ocorrência simultânea de A e B, é simplificada por:

$$\left.\begin{matrix} P(A|B) = P(A) \\ P(B|A) = P(B) \end{matrix}\right\} \Leftrightarrow \quad P(A \cap B) = P(A) \times P(B)$$

8

Sex \ Cor. fav.	Preta	Branca	Vermelha	Azul	Tot.
M	48	12	33	57	150
H	35	46	42	27	150
Total	83	58	75	84	300

Problema 1:

1.1. Qual é probabilidade de ser mulher?

1.2. Qual é probabilidade de não ser mulher?

1.3. Qual é probabilidade de uma pessoa ter cor favorita preta?

1.4. Qual é probabilidade de ter cor favorita preta e ser mulher?

1.5. Qual é probabilidade de uma pessoa ser mulher ou ter cor favorita preta?

1.6. Qual é probabilidade de ter cor favorita preta dado que a mulher?

1.7. Qual é probabilidade de ser mulher sabendo que tem cor favorita preta?

Solução:

1.1. 0,5 1.2. 0,5 1.3. 0,28 1.4. 0,16

1.5. 0,62 1.6. 0,32 1.7. 0,57

9

Problema 2: baseada na tabela,

2.1. Investigue se $P(A|B) = P(A)$ ou se $P(A \cap B) = P(A) \times P(B)$

2.2. Investigue se $P(A \cap B) = 0$ ou se $P(A \cup B) = P(A) + P(B)$

2.3. Comenta os resultados obtidos nas alíneas anteriores.

Exercício 1

1) Determine a probabilidade de não ter cor favorita preta e ser mulher
2) Determine a probabilidade de ser mulher e ter cor favorita preta ou azul

Solução: 1). 0,34 2). 0,35

Sex \ Cor. fav.	Preta	Branca	Vermelha	Azul	Tot.
M	48	12	33	57	150
H	35	46	42	27	150
Total	83	58	75	84	300

10

Exercício 2

Considere os dados de entrada e saída dos estrangeiros (H-homens, M-mulheres) em Díli no ano de 2013 através do aeroporto Díli, na seguinte tabela:

Mês	Entrada		Saída		Subtotal (entrada+saida)		Totais
	H	M	H	M	H	M	
Jan	5131	2996	4342	2360	9473	5356	**14829**
Feb	3522	1753	3619	1803			
Mar	3778	1928	4241	2258			
Abr	3991	2165	3966	1983			
Maio	4444	2209	4744	2593			
Jun	5014	3025	5161	2990			
Jul	5835	3453	6182	3473			
Agus	5169	2692	5754	3159			
Sept	5549	3186	5725	3120			
Out	5228	2627	2578	3031			
Nov	5723	2955	5831	2873			
Dez	5178	2968	5884	3485			
Totais	**58562**						

11

Complete os totais da tabela e determine:

1) Qual é probabilidade de um passageiro ser mulher?
2) Qual é probabilidade de um passageiro estar de saída?
3) Qual é probabilidade de um passageiro estar no aeroporto no mês de janeiro?
4) Qual é probabilidade de um passageiro ser mulher e sair do país?
5) Qual é probabilidade de um passageiro estar a entrar ou ser mulher?
6) Qual é probabilidade de um passageiro ser homem e entrar no mês de agosto?
7) Qual é probabilidade de um passageiro que saiu em 2013 ser mulher?
8) Qual é probabilidade de um passageiro que entrou não ser homem?
9) Qual é probabilidade de um passageiro que entra ser homem e entrar no mês de maio?
10) Qual é probabilidade de um passageiro de saída ser homem ou estar no aeroporto no mês de maio?

Solução: 1). 0,36 2). 0,50 3). 0,08 4). 0,18 5). 0,68
6). 0,03 7). 0,36 8). 0,35 9). 0,05 10). 0,67

12

2. Teste hipótese

2.1. Teste de independência de Qui-quadrado e de razão verossimilhanças

São os testes não paramétrico e que se aplica a amostra independentes. Simbolicamente denotado por χ^2 e G^2, respetivamente, para testar hipóteses estatísticas.

13

a) Teste de independência de Qui-quadrado

Objetivo: testar se A e B (variável linha e coluna numa tabela de contingência) são independentes.

Hipóteses a testar:

H_0 : as variáveis são independentes (não estão associadas).

H_1 : as variáveis não são independentes.

Assim, pretendem-se comparar as frequências observadas (O_{ij}) de cada uma das $r \times c$ células, com as correspondentes frequências esperadas (E_{ij}) supondo H_0 verdadeiro.

Frequências observadas (O_{ij})

		B	Totais
		$B_1 \ \ldots \ B_c$	
A	A_1	$O_{11} \ldots O_{1c}$	$O_{1.}$
	$\vdots$	$\vdots \quad \vdots$	$\vdots$
	A_r	$O_{r1} \ldots O_{rc}$	$O_{r.}$
Totais		$O_{.1} \ldots O_{.c}$	n

Comparar

Frequências esperadas (E_{ij})

		B	Totais
		$B_1 \ \ldots \ B_c$	
A	A_1	$E_{11} \ldots E_{1c}$	$E_{1.}$
	$\vdots$	$\vdots \quad \vdots$	$\vdots$
	A_r	$E_{r1} \ldots E_{rc}$	$E_{r.}$
Totais		$E_{.1} \ldots E_{.c}$	n

14

***Como obter* frequências esperadas (E_{ij})**

se H_0 for verdadeiro, isto é, as variáveis A e B forem independentes então:

$$E_{ij} = n \times P(A_i \cap B_j) \underset{\downarrow \ H_0 \text{ Verd.}}{=} n \times P(A_i) \times P(B_j) = n \times \frac{O_{i.}}{n} \times \frac{O_{.j}}{n} = \frac{O_{i.} \times O_{.j}}{n}$$

Pois a probabilidade de uma intersecção é igual ao produto das probabilidades.

Portanto, as hipóteses deste teste são:

$H_0: O_{ij} = E_{ij}, ; \forall_{ij}$

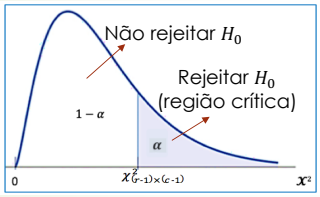
$H_1: O_{ij} \neq E_{ij}$; para algum i, j

15 *Estatística de teste*

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Onde: O_{ij} = frequências observadas
E_{ij} = frequências esperadas

- Quando o número de observações é elevado, a estatística χ^2 é aproximadamente a Qui-quadrado com $(r-1)\times(c-1)$ graus de liberdade (isto é, $\chi^2 \sim \boldsymbol{\chi^2_{(r-1)\times(c-1)}}$)



- A distribuição χ^2 é assimétrica à direita
- Os valores de χ^2 são positivos, ou seja, $\chi^2 \geq 0$.
- Um valor de χ^2 grande (pequeno) representa menor (maior) diferença entre O_{ij} e E_{ij}
- Teste unilateral à direita (i.e., região crítica à direita)

- Rejeita-se H_0 (em favor de H_1) quando χ^2_{obs} é maior ou igual ao valor crítico do teste, seja $\boldsymbol{\chi^2_{(1-\alpha);(r-1)\times(c-1)}}$. Se $\chi^2_{obs} < \boldsymbol{\chi^2_{(1-\alpha);(r-1)\times(c-1)}}$ então não se rejeita H_0.

16 Como calcular o valor crítico

Por exemplo, sendo: $r = 5$, $c = 6$ e nível de significância $\alpha = 5\%$.

O valor crítico:

$$\chi^2_{(1-\alpha);(r-1)\times(c-1)} = \chi^2_{(1-0,05);(5-1)\times(6-1)}$$
$$= \chi^2_{0,95;4\times5}$$
$$= \chi^2_{0,95;20}$$
$$= 31,410$$

Distribuição do Qui-Quadrado - χ^2_n

Os valores tabelados correspondem aos pontos x tais que: $P(\chi^2_n \leq x)$

	$P(\chi^2_n \leq x)$												
n	0,005	0,01	0,025	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,975	0,99	0,995
1	3,93E-05	0,000157	0,000982	0,003932	0,016	0,102	0,455	1,323	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,773	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	2,675	4,351	6,626	9,236	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	4,255	6,346	9,037	12,017	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	3,490	5,071	7,344	10,219	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	5,899	8,343	11,389	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342	12,549	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340	14,845	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,041	9,299	12,340	15,984	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	10,165	13,339	17,117	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	11,037	14,339	18,245	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	11,912	15,338	19,369	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338	20,489	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	13,675	17,338	21,605	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	14,562	18,338	22,718	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	15,452	19,337	23,828	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	16,344	20,337	24,935	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	17,240	21,337	26,039	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	18,137	22,337	27,141	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	19,037	23,337	28,241	33,196	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	16,473	19,939	24,337	29,339	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	20,843	25,336	30,435	35,563	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	18,114	21,749	26,336	31,528	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	22,657	27,336	32,620	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	23,567	28,336	33,711	39,087	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	20,599	24,478	29,336	34,800	40,256	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	29,051	33,660	39,335	45,616	51,805	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	37,689	42,942	49,335	56,334	63,167	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	46,459	52,294	59,335	66,981	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	61,698	69,334	77,577	85,527	90,531	95,023	100,425	104,215
80	51,172	53,540	57,153	60,391	64,278	71,145	79,334	88,130	96,578	101,879	106,629	112,329	116,321
90	59,196	61,754	65,647	69,126	73,291	80,625	89,334	98,650	107,565	113,145	118,136	124,116	128,299
100	67,328	70,065	74,222	77,929	82,358	90,133	99,334	109,141	118,498	124,342	129,561	135,807	140,170

17

Condições de aplicabilidade

- $n \geq 20$
- Todos os E_{ij} forem superiores a 1
- Pelo menos 80% dos E_{ij} forem não inferiores a 5

O teste do qui-quadrado é aplicável se todas as condições acima mencionadas são satisfazem. Caso contrário, a distribuição associada à estatística de teste é desconhecida, pelo que não se sabe como decidir sobre o H_0.

18

Exemplo

Retornem-se os dados do exemplo sobre as cores favoritas das pessoas.

Sex \ cor.fav.	Preta	Branca	Vermelha	Azul	Totais
M	48	12	33	57	**150**
H	35	46	42	27	**150**
Totais	**83**	**58**	**75**	**84**	**300**

Testar se a cor favorita de uma pessoa é dependente do género.

Resolução:

Hipóteses a testar:

H_0: a cor favorita de uma pessoa não depende do género

H_1: a cor favorita de uma pessoa depende do género

19

Procedimento do teste:

Tabela de frequências observadas (O_{ij})

Sex \ cor.fav.	Preta	Branca	Vermelha	Azul	Totais
M	48	12	33	57	**150**
H	35	46	42	27	**150**
Totais	**83**	**58**	**75**	**84**	**300**

Calcular as frequências esperadas: $E_{11} = \frac{O_{1.} \times O_{.1}}{n} = \frac{150 \times 83}{300} = 41{,}5$

$$\vdots$$

$$E_{24} = \frac{O_{2.} \times O_{.4}}{n} = \frac{150 \times 84}{300} = 42$$

Tabela de frequências esperadas (E_{ij})

Sex \ cor.fav.	preta	Branca	vermelha	Azul	Totais
M	41,5	29	37,5	42	**150**
H	41,5	29	37,5	42	**150**
Totais	**83**	**58**	**75**	**84**	**300**

20

Verificar as condições necessários

- A amostra é grande ($n = 300 > 20$)
- Todas as frequências esperadas são superiores a 5

Teste estatística

$$\chi^2_{obs} = \frac{(48-41{,}5)^2}{41{,}5} + \frac{(35-41{,}5)^2}{41{,}5} + \frac{(12-29)^2}{29} + \frac{(46-29)^2}{29} + \frac{(33-37{,}5)^2}{37{,}5} + \frac{(42-37{,}5)^2}{37{,}5} + \frac{(57-42)^2}{42} + \frac{(27-42)^2}{42}$$

$$\chi^2_{obs} \approx \mathbf{33{,}76}$$

Nível significância

$\alpha = 0{,}05$

Valor crítico

$\chi^2_{(1-\alpha);(r-1)\times(c-1)} = \chi^2_{(1-0{,}05);(2-1)\times(4-1)} = \boldsymbol{\chi^2_{0{,}95;3}} = \mathbf{7{,}82}$

Conclusão

Como $\boldsymbol{\chi^2_{obs} > \chi^2_{0{,}95;3}}$, rejeita-se H_0 em favor de H_1. Conclui-se, ao nível significância 5%, que a cor favorita de uma pessoa depende do género.

21

b) Teste de razão verosimilhanças

O teste de razão verosimilhanças é um teste não paramétrico que se aplica a uma amostra independente representada por uma tabela de contingência. É semelhante ao teste X^2 mas apresenta uma estatística de teste diferente.

		$\boldsymbol{B}$	Tot.
		$B_1 \ldots B_c$	
$\boldsymbol{A}$	A_1	$O_{11} \ldots O_{1c}$	$\boldsymbol{O_{1.}}$
	.	. .	.
	.	. .	.
	A_r	$O_{r1} \ldots O_{rc}$	$\boldsymbol{O_{r.}}$
Totais		$\boldsymbol{O_{.1} \ldots O_{.c}}$	$\boldsymbol{n}$

Hipóteses a testar

H_0: as variáveis são independentes

H_1: $\sim H_0$

Se H_0 for verdadeira, as frequências esperadas são estimadas por:

$$E_{ij} = \frac{O_{i.} \times O_{.j}}{n}$$

Estatística de teste

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c}\left(O_{ij}\ln\left(\frac{O_{ij}}{E_{ij}}\right)\right) \sim \chi^2_{(r-1)\times(c-1)}$$

Onde: O_{ij} = frequências observadas

E_{ij} = frequências esperadas

Decisão: Rejeita-se H_0 (em favor de H_1) quando $G^2 \geq \chi^2_{(1-\alpha);(r-1)\times(c-1)}$. Caso contrário não se rejeita H_0.

22

Exemplo

Retornem-se os dados do exemplo sobre as cores favoritas das pessoas.

Tabela de frequências observadas (O_{ij})

	Pre	Bra.	Ver.	Az.	Tot.
M	48	12	33	57	**150**
H	35	46	42	27	**150**
Tot.	**83**	**58**	**75**	**84**	**300**

Tabela de frequências esperadas (E_{ij})

	pre	Bra.	Ver.	Az.	Tot.
M	41,5	29	37,5	42	**150**
H	41,5	29	37,5	42	**150**
Tot.	**83**	**58**	**75**	**84**	**300**

Teste estatística

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c}\left(O_{ij}\ln\left(\frac{O_{ij}}{E_{ij}}\right)\right) = 2\times\left[\left(48\times\ln\left(\frac{48}{41{,}5}\right)\right) + \cdots + \left(27\times\ln\left(\frac{27}{42}\right)\right)\right] = 35{,}35$$

Valor crítico

$\chi^2_{(1-\alpha);(r-1)\times(c-1)} = \chi^2_{0{,}95;3} = 7{,}82$

Conclusão: Como $\boldsymbol{G^2 > \chi^2_{0{,}95;3}}$, rejeita-se H_0 (em favor de H_1). Conclui-se, ao nível significância 5%, que a cor favorita de uma pessoa é depende do género.

23

c) Medidas de associação

Objetivo: medir o nível de associação entre duas variáveis consideradas numa tabela de contingência $r \times c$.

		$\boldsymbol{B}$	Totais
		$B_1 \ \ldots \ B_c$	
$\boldsymbol{A}$	A_1	$O_{11} \ldots O_{1c}$	$O_{1.}$
	⋮	⋮ ⋮	⋮
	A_r	$O_{r1} \ldots O_{rc}$	$O_{r.}$
Totais		$O_{.1} \ldots O_{.c}$	n

Teste de independência

H_0: as variáveis são independentes (não estão associadas).

H_1: $\sim H_0$

Em caso de rejeitar H_0, como quantificar o nível de associação? $\Longrightarrow$ **Medidas de associação** baseadas no valor de χ^2_{obs}.

24

Medidas de associação:

- Coeficiente de contingência de Pearson
- Coeficiente de Tschuprow
- Coeficiente de contingência de V de Cramér
- Coeficiente fi

Interpretação:

- Se $\chi^2_{obs} = 0$ então a medida de associação deve ser igual à zero (situação de independência).
- Quanto maior for a medida de associação, maior o grau de dependência entre as duas variáveis.

25

Seja r o número de linhas e c o número de colunas da tabela de contingência.

- **Coeficiente de contingência de Pearson, C**

$$C = \sqrt{\frac{\chi^2_{obs}}{\chi^2_{obs} + n}}$$

Limite de variação: $0 \leq C \leq \sqrt{\frac{min\{r,c\}-1}{min\{r,c\}}}$, onde o limite superior é sempre inferior a 1.

- **Coeficiente de Tschuprow, T**

$$T = \sqrt{\frac{\chi^2_{obs}}{n\sqrt{(r-1)(c-1)}}}$$

Limite de variação: $0 \leq T \leq 1$, onde $T = 1$ ocorre apenas para $r = c$.

- **Coeficiente de contingência de V de Cramér, V**

$$V = \sqrt{\frac{\chi^2_{obs}}{n[min\{r,c\} - 1]}}$$

Limite de variação: $0 \leq V \leq 1$

NOTA:
Em geral, $V \geq T$ exceptuando para $r = c$ em que $V = T$

26

- **Coeficiente fi, ϕ para $r = c = 2$**

Supondo H_0 verdadeiro, a estatística de teste de χ^2 é

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Logo, $\varphi = \sqrt{\frac{\chi^2_{obs}}{n}} = \frac{|ad-bc|}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$, $0 \leq \varphi \leq 1$

		Variável x_1		Total
		1	2	
Variável x_2	1	a	b	$a+b$
	2	c	d	$c+d$
Total		$a+c$	$b+d$	n

Alternativamente, $\varphi' = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$; $-1 \leq \varphi' \leq 1$

Neste caso, mede-se a intensidade da associação e também a sua direção

- $\varphi = 0 \rightarrow ad - bc = 0 \Longrightarrow$ ausência de associação (i.e., independência)
- $\varphi > 0 \rightarrow ad > bc \Longrightarrow$ associação positiva
- $\varphi < 0 \rightarrow ad < bc \Longrightarrow$ associação negativa

27 **Exemplo**

1. Retornem-se os dados do exemplo sobre as cores favoritas das pessoas (do teste de Qui-quadrado).

Como já viu $\chi^2_{obs} \approx 33{,}76$, com $\chi^2_{(2-1)\times(4-1)} = \chi^2_3$ que implica rejeitar H_0 de independência.

As medidas de associação para este exemplo são as seguintes:

$$C = \sqrt{\frac{\chi^2_{obs}}{\chi^2_{obs}+n}} \approx \sqrt{\frac{33{,}76}{33{,}76+300}} \approx 0{,}32, \text{ onde } 0 \leq C \leq \sqrt{\frac{min\{2,4\}-1}{min\{2,4\}}} = \sqrt{\frac{1}{2}} \approx 0.71$$

$$T = \sqrt{\frac{\chi^2_{obs}}{n\sqrt{(r-1)(c-1)}}} \approx \sqrt{\frac{33{,}76}{300\times\sqrt{(2-1)(4-1)}}} = \sqrt{\frac{33{,}76}{300\times\sqrt{3}}} \approx 0{,}25, \text{ onde } 0 \leq T \leq 1$$

$$V = \sqrt{\frac{\chi^2_{obs}}{n[min\{r,c\}-1]}} \approx \sqrt{\frac{33{,}76}{300\times[min\{2,4\}-1]}} = \sqrt{\frac{33{,}76}{300\times1}} \approx 0{,}34, \text{ onde } 0 \leq T \leq 1$$

∴ A associação entre a cor preferida e o género é moderada/forte (aproximação ao limite superior).

Existem evidências de que a associação não é fraca (aproximação a zero).

28 2. Testar se há associação entre tabagismo e a prática de desporte, e quantificar o seu grau de associação, se existir.

		Tabagismo		Total
		Presente	Ausente	
Prática de desporto	Presente	50	15	**65**
	Ausente	10	25	**35**
Total		**60**	**40**	**100**

Hipóteses a testar:

H_0: não há associação entre tabagismo e prática desportiva

H_1: $\sim H_0$

Valor observado da estatística de teste :

$$\chi^2_{obs} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{100\times(50\times25-15\times10)^2}{(50+15)\times(10+25)\times(50+10)\times(15+25)} \approx 22{,}16$$

29

Valor crítico:

$$\alpha = 0{,}05 \rightarrow \chi^2_{(1-\alpha);(r-1)\times(c-1)} = \chi^2_{(1-0{,}05);(2-1)\times(2-1)} = \chi^2_{0{,}95;1} = 3{,}84$$

Decisão:

Como $\chi^2_{obs} > \chi^2_1$ rejeita-se H_0 em favor de H_1. Conclui-se, ao nível significância 5%, que existe uma associação significativa entre tabagismo e prática desportiva.

Como H_0 é rejeitada, logo, o nível de associação entre as variáveis deve ser quantificado. Assim,

$$\varphi = \sqrt{\frac{\chi^2_{obs}}{n}} \approx \sqrt{\frac{22{,}16}{100}} \approx 0{,}47, \text{ onde } 0 \leq \varphi \leq 1$$

∴ A associação entre o tabagismo e a prática desportiva é moderada/forte. Além disso associação é positiva, isto é, presença/ausência de prática desportiva é associada com presença/ausência de tabagismo.

30

2.2. Outro teste de Qui-quadrado

Além de teste de Qui-quadrado de independência, existe outros testes de Qui-quadrado que podem ser aplicados a tabelas de contingência, são

- Teste de homogeneidade
- Teste de ajustamento ou teste de aderência

Embora testem hipóteses diferentes, ambos os testes de hipóteses têm cálculos semelhantes aos efetuados para o teste de Qui-quadrado de independência.

31

a) Teste de homogeneidade de Qui-quadrado

Objetivo: testar se as populações em A são homogéneas.

Hipóteses a testar

H_0: as proporções em $A_1, \dots, A_r$ são iguais para todas as categorias de B (isto é, as populações são homogéneas).

H_1: $\sim H_0$

Frequências observadas (O_{ij})

		B	
		$B_1 \dots B_c$	Totais
A	A_1	$O_{11} \dots O_{1c}$	$O_{1.}$
	A_r	$O_{r1} \dots O_{rc}$	$O_{r.}$
Totais		$O_{.1} \dots O_{.c}$	n

Comparar

Frequências esperadas (E_{ij})

		B	
		$B_1 \dots B_c$	Totais
A	A_1	$E_{11} \dots E_{1c}$	$E_{1.}$
	A_r	$E_{r1} \dots E_{rc}$	$E_{r.}$
Totais		$E_{.1} \dots E_{.c}$	n

♯ esperado de indivíduos na intersecção de A_1 e B_1

Os teste de homogeneidade e o teste de independência distinguem-se pela forma de como as amostras são recolhidas. Tipicamente, num teste de homogeneidade fixam-se os totais marginais para populações.

32

Como funciona o teste de homogeneidade

A realização deste teste é semelhante à do teste de independência, isto é, as hipóteses deste teste são equivalente às do teste de independência. Isto é,

$H_0: O_{ij} = E_{ij}; \forall_{ij}$

$H_1: O_{ij} \neq E_{ij}$; para algum i, j

Estatística de teste: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \boldsymbol{\chi^2_{(r-1)\times(c-1)}}$

Decisão: Rejeita-se H_0 (em favor de H_1) quando $\chi^2_{obs} \geq \boldsymbol{\chi^2_{(1-\alpha);(r-1)\times(c-1)}}$. Senão, não se rejeita H_0 ao nível de significância α.

33 **Exemplo**

Retornem-se os dados do exemplo sobre as cores favoritas das pessoas. Testar se os géneros mulher e homem têm as mesmas proporções.

Sex \ cor.fav.	Preta	Branca	Vermelha	Azul	Totais
M	48	12	33	57	**150**
H	35	46	42	27	**150**
Totais	**83**	**58**	**75**	**84**	**300**

Nota:

- Ao fixar os totais para os grupos M e H, o teste de homogeneidade deverá comparar M e H.
- Se o objetivo do estudo fosse testar se as cores são escolhidas de forma homogénea então a recolha de dados devia fixar o número de cores escolhidas.

34 **Resolução:**

Hipóteses a testar:

H_0: a mulher e o homem têm os mesmos proporções

H_1: a mulher e o homem têm diferentes proporções

Procedimento do teste:

É semelhante à do teste de independência. Isto é, as frequências esperadas, o valor de $\boldsymbol{\chi^2_{obs}}$ e o valor crítico (em nível de significância 5%) são calculados tal como no exemplo do teste de independência.

Assim, temos: $\boldsymbol{\chi^2_{obs} \approx 33,76}$ e $\boldsymbol{\chi^2_{0,95;3} = 7,82}$

Conclusão

Como $\boldsymbol{\chi^2_{obs} > \chi^2_{0,95;3}}$, logo H_0 é rejeitada em nível significância 5% . Ou seja, ao nível significância 5%, o homem e a mulher têm diferentes proporção em relação das suas cores favoritas.

35

b) Teste de ajustamento (teste de aderência) de Qui-quadrado

Objetivo: testar se as observações seguem uma determinada distribuição teórica de probabilidade (discreta ou contínua, com ou sem parâmetros conhecidos).

Hipóteses a testar

H_0: A população segue uma determinada distribuição $\mathcal{D}$

H_1: A população não tem uma determinada distribuição $\mathcal{D}$

	C_1 ... C_k	Total
Freq.obs.	O_1 ... O_k	n
Freq.esp.	E_1 ... E_k	n

36

Como funciona o teste de ajustamento

Para a realização do teste, os dados tem que estar agrupados em k classes (intervalos ou categorias). No caso em que a distribuição $\mathcal{D}$ é contínua, tais classes podem ser baseadas nas classes do histograma.

Neste teste também são comparadas duas quantidades:

- O número de observações em cada categoria (frequência observada, O_i)
- O número de valores que se teriam em cada categoria admitindo que a população tem a distribuição $\mathcal{D}$ (frequência esperada, E_i).

Assim, as hipóteses deste teste são:

$H_0: O_i = E_i \; ; \forall_i$

$H_1: O_i \neq E_i$; para algum i

A frequência esperada de uma classe, quando H_0 é verdadeira, é dada por:

$E_i = n \times P_i$ com $P_i = P(C_i)$

37

Estatística de teste

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \boldsymbol{\chi^2_{(1-\alpha);(k-m-1)}}$$

Quando o número de observações é elevado, a estatística χ^2 segue aproximadamente uma distribuição Qui-quadrado com $(k-m-1)$ graus de liberdade, onde:

- k representa o número de categorias e
- m representa o número de parâmetros de $\mathcal{D}$ que é necessário estimar a partir da amostra.

Decisão : Rejeita-se H_0 (em favor de H_1) quando $\chi^2_{obs} \geq \boldsymbol{\chi^2_{(1-\alpha);(k-m-1)}}$. Senão, não se rejeita H_0 ao nível de significância α.

38

Exemplo

Retornem-se os dados do exemplo sobre as cores favoritas das pessoas.

Sex \ cor.fav.	Preta	Branca	Vermelha	Azul	Totais
M	48	12	33	57	**150**
H	35	46	42	27	**150**
Totais	**83**	**58**	**75**	**84**	**300**

Testar se a distribuição das cores favoritas nas mulheres é uniforme.

Resolução:

Hipóteses a testar:

H_0: a distribuição da probabilidade de mulher tem uma distribuição uniforme

H_1: a distribuição da probabilidade de mulher tem outra distribuição

39

Procedimento do teste:

Probabilidade de cada categoria

Como a distribuição é uniforme seria um número finito de resultados com chances (possibilidades) iguais de acontecer. Logo, neste caso os $P_i's$ são iguais, i.e., $1/4$.

Tabela de frequências observadas e frequências esperadas de mulheres:

	Preta	Branca	Vermelha	Azul	Total
$\boldsymbol{O_i}$	48	12	33	57	**150**
$\boldsymbol{P_i}$	1/4	1/4	1/4	1/4	**1**
$\boldsymbol{E_i = n \times P_i}$	37,5	37,5	37,5	37,5	**150**

Outra maneira de calcular o número de esperado:

$$E_i = \frac{n}{k} = \frac{150}{4} = 37{,}5$$

40

Estatística de teste

$$\chi^2_{obs} = \frac{(48-37{,}5)^2}{37{,}5} + \frac{(12-37{,}5)^2}{37{,}5} + \frac{(33-37{,}5)^2}{37{,}5} + \frac{(57-37{,}5)^2}{37{,}5} \approx \mathbf{30,96}$$

Nível significância

$\alpha = 0{,}05$

Valor crítico

$m = 0, k = 4$. Logo, $\boldsymbol{\chi^2_{(1-\alpha);(k-m-1)} = \chi^2_{0{,}95;3} = 7{,}82}$

Conclusão

Como $\boldsymbol{\chi^2_{obs} > \chi^2_{0{,}95;3}}$, conclui-se de rejeitar H_0. Isto é, em nível de significância 5%, a distribuição de mulheres em relação às cores favoritas não se ajusta a uma distribuição uniforme.

41 **Exercício 1**

1. Identifique o teste adequado para os seguintes problemas:

a) Avaliar se existe associação significativa entre a cor de uma flor de uma espécie de planta e a existência de um tipo de parasita. Avaliaram-se vários exemplares dessa espécie e registou-se a cor e a existência ou não do parasita.

b) Avaliar o crescimento de uma planta (baixo, médio e alto) face à presença de luz solar direta e indireta. Submete-se durante um ano 50 exemplares à exposição diária direta e 50 exemplares à exposição indireta.

c) Em 100 lançamento de um dado, pretende-se avaliar que os resultados obtidos sustentam que o dado é honesto.

Solução:

a) Teste de independência

b) Teste de homogeneidade (o crescimento não depende do tipo de exposição solar)

c) Teste de ajustamento

42 **Exercício 2** – Identifica qual é o teste adequada para os seguintes problemas e resolver:

2.1. Uma pesquisa sobre a qualidade de um serviço público foi realizada enviando-se questionários pelo correio com porte pago. Desconfiando-se que poderia haver um viés nas respostas, fez-se também uma pesquisa por e-mail e outra por telefone. Os resultados estão abaixo. Há relação entre a forma de pesquisa e os seus resultados?

	Correio	E-mail	Telefone	Totais
Excelente	62	36	24	**122**
Satisfatório	48	42	16	**106**
Insatisfatório	24	22	20	**66**
Totais	**134**	**100**	**60**	**294**

Solução:

Teste adequado é teste de independência

Resultado: $\chi^2_{obs} = 8{,}35$ e $\chi^2_{0{,}95;4} = 9{,}49$, logo não se rejeitar o H_0 em favor de H_1 isto é opinião é independente da forma de pesquisa.

43

2.2. Existem três modos de efetuar os pagamentos num supermercado durante o período do dia, são: por cheque, dinheiro e cartão de crédito. A seguinte tabela de contingência apresenta os resultados obtidos numa amostra de 4000 clientes:

	Período do dia		
Modo de pagamento	Manhã	Tarde	Noite
Cheque	750	1500	750
Dinheiro	125	300	75
Cartão de crédito	125	200	175

Testar ao nível de significância de 5% a hipótese de que os três modos de pagamentos dos clientes tem mesma proporção em relação ao período do dia em que fazem as compras.

Solução: Teste adequado é teste de homogeneidade

Resultado: $\boldsymbol{\chi^2_{obs}} = 60$ e $\chi^2_{0,95;4} = 9{,}49$, logo concluímos rejeitar H_0 em favor de H_1 e portanto os três modos de pagamento tem proporções diferentes em relação ao período dia.

44

2.3. Pensa-se que o número de defeitos por circuito, num certo tipo de circuitos, deve seguir uma distribuição de Poisson. De uma amostra (escolhida aleatoriamente) de 60 circuitos obtiveram-se os resultados seguintes:

Nº. de def.	O_i
0	32
1	15
2	9
3	4
Total	**60**

Solução:

$\boldsymbol{\chi^2_{obs} \approx 2{,}96}$ e $\boldsymbol{\chi^2_{(1-\alpha);(k-m-1)} = \chi^2_{0{,}95;\,1} = 3{,}84}$, logo não rejeitar H_0 em favor de H_1

45

2.3. Teste alternativa de independência para tabelas 2×2

Tabela contingência $r = c = 2$

		Variável x_1		Total
		1	2	
Variável x_2	1	a	b	$a+b$
	2	c	d	$c+d$
Total		$a+c$	$b+d$	n

Alternativa ao teste do Qui-quadrado

- **Teste exato de Fisher**

  Para amostras independentes e quando o teste Qui-quadrado não se aplica, isto é,

  - Para $n < 20$ ou
  - Para $20 < n < 40$ e existe pelo menos uma $E_{ij} < 5$

- **Teste McNemar**

  Para amostras emparelhadas ou correlacionadas.

46

a) Teste exato de Fisher

É chamado exato, porque, permite a calcular a probabilidade exata de ocorrência de uma frequência observada, ou de valor mais extremos (P_value).

No teste independência, as hipóteses a testar são:

H_0: independência

H_1: $\sim H_0$

Se H_0 é verdadeira, então $E_{ij} = \frac{O_i \times O_j}{n}$; $i = \{1, 2\}$, $j = \{1, 2\}$

47

- Se $O_{11} > E_{11}$ então a associação é positiva (cauda da direita)

H_0: independência

H_1: associação positiva

		Variável x_1		Total
		1	2	
Variável x_2	1	$\boldsymbol{a}$	b	$a+b$
	2	c	$\boldsymbol{d}$	$c+d$
Total		$a+c$	$b+d$	n

Neste caso, O_{11} é superior ao esperado E_{11} (segundo independência) e portanto há mais observações na diagonal principal.

- Se $O_{11} < E_{11}$ então a associação é negativa (cauda da esquerda)

H_0: independência

H_1: associação negativa

		Variável x_1		Total
		1	2	
Variável x_2	1	a	$\boldsymbol{b}$	$a+b$
	2	$\boldsymbol{c}$	d	$c+d$
Total		$a+c$	$b+d$	n

Neste caso, O_{11} é inferior ao esperado E_{11} (segundo independência) e portanto há mais observações na diagonal secundaria.

48

Cálculo do $P_{\_Value}$

		Var.x_1		Tot.
		1	2	
Var. x_2	1	a	b	$a+b$
	2	c	d	$c+d$
Tot.		$a+c$	$b+d$	n

$$P_a = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$$

P_a = probabilidade de observar $O_{11} = a$, mantendo os totais

- H_1: associação positiva (cauda da direita)

$$\overset{a}{\widetilde{O_{11}}} > E_{11} = \frac{\overset{a+b}{\widetilde{O_{1.}}} \times \overset{a+c}{\widetilde{O_{.1}}}}{n}$$

$$\therefore P_{\_Value} = P_a + P_{a+1} + \cdots + P_{min(a+b,a+c)} = P_{sup}$$

Probabilidade de observar $O_{11} = a$ ou superior

- H_1: associação negativa (cauda da esquerda)

$$\overset{a}{\widetilde{O_{11}}} < E_{11} = \frac{\overset{a+b}{\widetilde{O_{1.}}} \times \overset{a+c}{\widetilde{O_{.1}}}}{n}$$

$$\therefore P_{\_Value} = P_a + P_{a-1} + \cdots + P_0 = P_{inf}$$

Probabilidade de observar $O_{11} = a$ ou inferior

49 **Regra de decisão**

H_0: independência

Unilateral

H_1: associação positiva ou H_1: associação negativa

Regra de unilateral: *Se* $P_{\_Value} \leq \alpha$ rejeitar H_0 em favor de H_1, senão (i.e., $P_{\_Value} \geq \alpha$) não rejeitar H_0

Bilateral:

H_1: existe associação

Regra de bilateral: Se $P_{sup} \leq \frac{\alpha}{2}$ ou $P_{inf} \leq \frac{\alpha}{2}$ rejeitar H_0 em favor de H_1

50 **Exemplo**

1. Relativamente ao aparecimento de determinada doença, obtiveram-se os seguintes dados, numa amostra de 9 pessoas

	Mulher	Homem	**Totais**
Doentes	1	3	**4**
Não doentes	3	2	**5**
Totais	**4**	**5**	**9**

Pretende-se testar ($\alpha = 0{,}05$) se:

a) Existe uma associação entre o aparecimento de determinada doença e o género

b) A proporção de doentes é diferente nos homens e nas mulheres

51

Resolução do exemplo:

1. a). $O_{11} = 1$ e $E_{11} = \frac{O_{1.} \times O_{.1}}{n} = \frac{4 \times 4}{9} = \frac{16}{9} = 1{,}78 \Longrightarrow \boldsymbol{O_{11} < E_{11}}$

Hipóteses a testar:

H_0: não existe associação entre a aparecimento de determinada doença e o género

H_1: existe uma associação negativa entre as variáveis

	Mulher	Homem	**Totais**
Doentes	1	3	**4**
Não doentes	3	2	**5**
Totais	**4**	**5**	**9**

$P_1 = \frac{4!5!4!5!}{9!1!3!3!2!} \approx 0{,}317$

	Mulher	Homem	**Totais**
Doentes	0	4	**4**
Não doentes	4	1	**5**
Totais	**4**	**5**	**9**

$P_0 = \frac{4!5!4!5!}{9!0!4!4!1!} \approx 0{,}040$

Assim a probabilidade de significância $P_{\_Value} = P_1 + P_0 \approx 0{,}317 + 0{,}040 \approx 0{,}357$.

Como $0{,}357 > 0{,}05$, não se rejeita H_0 em favor de H_1 e conclui-se que não existe associação significativa entre o aparecimento de determinada doença e o género.

52

b). *Hipóteses a testar:*

H_0: não existe diferença na proporção de doentes entre mulheres e homens

H_1: $\sim H_0$

Na alínea a), obteve-se $P_{\_Value} \approx 0{,}357$.

Uma vez que $P_{\_Value} > \frac{\alpha}{2}$ (teste bilateral), seja, $0{,}357 > 0{,}025$ conclui-se não se rejeita H_0 em favor de H_1 e assim para um nível de significância de 5%, a proporção de doentes não difere significativamente entre mulheres e homens.

53

Exercício

Considere-se o resultado de um estudo feito para comparar a eficácia de dois tratamentos e em que 7 pacientes receberam o tratamento I e 8 pacientes o tratamento II.

Resultados segundo o tratamento			
Classificação	Tratamento I	Tratamento II	Totais
Curados	4	1	**5**
Não curados	3	7	**10**
Totais	**7**	**8**	**15**

Testar se existe associação entre o resultado e o tipo de tratamento ($\alpha = 0{,}01$).

Solução:

$P_{\_Value} \approx 0{,}10$

Decisão: não se rejeita H_0 em nível de significância 1%, e assim não existe a associação significativa entre o resultado e o tipo de tratamento.

54

b) Teste de McNemar

Objetivo: avaliar a eficiência de situações "antes" e "depois".

- $\boldsymbol{a}$ e $\boldsymbol{d}$ é o número de indivíduos que não mudaram de condição
- $\boldsymbol{b}$ é o número de **insucessos** ⟶ indivíduos que mudaram de (+) para (-)
- $\boldsymbol{c}$ é o número de **sucessos** ⟶ indivíduos que mudaram de (-) para (+)
- $\boldsymbol{b+c}$ é o total de indivíduos que mudaram de condição.
- Condição exigida: $b + c > 10$

		Depois		Totais
		+	-	
Antes	+	a	b	$a+b$
	-	c	d	$c+d$
Totais		$a+c$	$b+d$	n

55

Hipóteses a testar:

H_0: não existe diferença antes e depois do tratamento

H_1: existe diferença antes e depois do tratamento

- Se $b + c \leq 20$, aplica-se o teste binomial

  $P_{\_Value} = P[X = x] = \binom{n}{x} P^x (1-P)^{n-x}$ onde $X \sim B(b + c, \frac{1}{2})$, com $x = min\{b, c\}$.

- Se $b + c > 20$, usa-se o teste de χ^2

$$\chi^2 = \frac{(b-c)^2}{b+c} \sim \chi^2_{(1)}$$

E assim, $P_{\_Value} = P(\chi^2_{(1)} > \chi^2_{obs})$

Decisão do teste:

Se $P_{\_Value} \leq \alpha$, rejeitar H_0 em favor de H_1

Senão (i.e., $P_{\_Value} > \alpha$) não se rejeitar H_0 em favor de H_1

56

Exemplo

Uma empresa, que realizou uma campanha publicitária para promoção de um produto em duas cidades diferentes, pretende saber se as preferências dos consumidores se modificaram após a dita campanha. As respostas de 80 consumidores na cidade I, e 100 consumidores na cidade II, apresentam-se nas seguintes tabelas:

Cidade I

Consumo		Depois		Total
		Sim	Não	
Antes	Sim	37	3	40
	Não	13	47	60
Total		50	50	100

Cidade II

Consumo		Depois		Total
		Sim	Não	
Antes	Sim	14	7	21
	Não	26	33	59
Total		40	40	80

Terá existido uma mudança significativa no consumo do produto após a campanha publicitária ($\alpha = 0{,}05$)?

57

Resolução do exemplo

– Cidade I

Consumo		Depois		Total
		Sim	Não	
Antes	Sim	37	3	40
	Não	13	47	60
Total		50	50	100

Hipóteses a testa:

$H_0: P(sim \rightarrow não) = P(não \rightarrow sim)$

$H_1: P(sim \rightarrow não) \neq P(não \rightarrow sim)$

Total de indivíduos que mudaram de condição é $3 + 13 = 16 < 20$, logo aplicar o teste binomial com $n = 16$ e $x = 3$.

$$P_{\_Value} = P[X = 3] = \binom{16}{3} \times \left(\frac{1}{2}\right)^3 \times \left(1 - \frac{1}{2}\right)^{16-3} = \frac{16!}{3!\,13!} \times \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{2}\right)^{13} \approx 0{,}01$$

Conclusão: como $P_{\_Value} < \alpha$, conclui-se rejeitar H_0, i.e., que a campanha publicitária teve influência no consumo do produto (para um nível de significância de 5%).

58

– Cidade II

Consumo		Depois		Total
		Sim	Não	
Antes	Sim	14	7	21
	Não	26	33	59
Total		40	40	80

Hipóteses a testa:

$H_0: P(sim \rightarrow não) = P(não \rightarrow sim)$

$H_1: P(sim \rightarrow não) \neq P(não \rightarrow sim)$

Total de indivíduos que mudaram de condição é $7 + 26 = 33 > 20$, logo usar o estatística de teste:

$$\chi^2_{obs} = \frac{(b-c)^2}{b+c} = \frac{(7-26)^2}{7+26} \approx 10{,}94$$

$$P_{\_Value} = P\left(\chi^2_{(1)} > \chi^2_{obs}\right) = P\left(\chi^2_{(1)} > 10{,}94\right) < 0{,}005$$

Conclusão: como $P_{\_Value} < \alpha$, conclui-se rejeitar o H_0, i.e., que a campanha publicitária teve influência no consumo do produto (para um nível de significância de 0,05).

59

Como H_0 é rejeitada, deve ser quantificar o nível de associação entre as variáveis. Assim,

$$\varphi = \sqrt{\frac{\chi^2_{obs}}{n}} \approx \sqrt{\frac{10{,}94}{80}} \approx 0{,}37, \text{ onde } 0 \leq \varphi \leq 1$$

∴ A associação entre a campanha publicitária e o preferências dos consumo do produto é moderada/forte (aproximação ao limite superior).

Existem evidências de que a associação não é fraca (aproximação a zero).

60 Foi uma boa influencia?

É preciso construir φ' a custa de φ de forma que:

- $\varphi' > 0 \implies$ efeito positivo
- $\varphi' < 0 \implies$ efeito negativo

Pois c é o número de sucessos e b é o número de insucessos

Se $c > b$ quer dizer que número de sucesso é maior do que número de insucesso, logo é uma boa influência.

$$\varphi = \sqrt{\frac{\chi^2_{obs}}{n}} = \sqrt{\frac{(b-c)^2}{n(b+c)}} = \frac{|b-c|}{\sqrt{n(b+c)}} \Rightarrow \varphi' = \frac{c-b}{\sqrt{n \times (c+b)}}$$

Assim,

$$\varphi' = \frac{26-7}{\sqrt{80 \times (26+7)}} \approx 0{,}37, \text{ onde } -1 \leq \varphi' \leq 1$$

∴ Existe evidências de que a campanha publicitária teve influência positiva no consumo do produto

61

Exercício

Foram inquiridos no âmbito de dois estados (um encomendando por um jornal diário e por um jornal um semanário) sobre a preferência entre o partido do governo e o maior partido da oposição, antes e depois de um importante debate entre os respetivos líderes. Os resultados encontraram-se sumariados nos quadros seguintes:

(i) Resultados do jornal diário

		Depois		Tot.
		Gov.	Op.	
Antes	Gov.	20	12	**32**
	Op.	8	15	**23**
Tot.		**28**	**27**	**55**

(ii) Resultados do jornal semanário

		Depois		Tot.
		Gov.	Op.	
Antes	Gov.	10	22	**32**
	Op.	10	13	**23**
Tot.		**20**	**35**	**55**

Testar se o debate teve influência na opinião dos leitores, com nível de significância 5%

Solução:

(i). $P_{\_Value} = 0{,}12$ (ii). $0{,}025 < P_{\_Value} < 0.05$

62

3. Tabelas de contingência $r \times c$

3.1. Localização de fontes de dependência por análise dos resíduos

No caso de rejeitar H_0 de independência, como identificar as células que mais contribuem para a dependência?

A localização destas fontes de dependência pode ser feita pela "**análise dos resíduos**".

63 Como calcular o resíduo (R_{ij})

O resíduo padronizado é dado por:

$$R_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}\left(1 - \frac{O_{i.}}{n}\right)\left(1 - \frac{O_{j}}{n}\right)}}$$

Frequências observadas (O_{ij})

		B		Totais
		B_1	B_c	
A	A_1	O_{11}	O_{1c}	$O_{1.}$
	A_r	O_{r1}	O_{rc}	$O_{r.}$
Totais		$O_{.1}$	$O_{.c}$	n

Frequências esperadas (E_{ij})

		B		Totais
		B_1	B_c	
A	A_1	E_{11}	E_{1c}	$E_{1.}$
	A_r	E_{r1}	E_{rc}	$E_{r.}$
Totais		$E_{.1}$	$E_{.c}$	n

Resíduos (R_{ij})

		B		Totais
		B_1	B_c	
A	A_1	R_{11}	R_{1c}	$R_{1.}$
	A_r	R_{r1}	R_{rc}	$R_{r.}$
Totais		$R_{.1}$	$R_{.c}$	n

Se H_0 de independência for verdadeira, então $R_{ij} \sim N(0,1)$ quando $n \to \infty$.

Comparando $|R_{ij}|$ com o quantil de probabilidade $Z_{1-\frac{\alpha}{2}}$ da distribuição normal reduzida:

- As células tais que $|R_{ij}| \geq Z_{1-\frac{\alpha}{2}}$, contribuem (de forma significativa) para a dependência das variáveis,
- Quanto maior for o valor de $|R_{ij}|$ maior é a contribuição para dependência.

64 Como determinar o quantil de probabilidade $Z_{1-\frac{\alpha}{2}}$

Sendo $\alpha = 0{,}05$

Então, $\frac{\alpha}{2} = 0{,}025$

Logo,

$$Z_{1-\frac{\alpha}{2}} = Z_{1-0{,}025} = Z_{0{,}975} = 1{,}96$$

Tabela da Distribuição Normal Reduzida

z	Φ(-z)	Φ(z)	z	Φ(-z)	Φ(z)	z	Φ(-z)	Φ(z)	z	Φ(-z)	Φ(z)	z	Φ(-z)	Φ(z)	z	Φ(-z)	Φ(z)
0.01	0.4960	0.5040	0.51	0.3050	0.6950	1.01	0.1562	0.8438	1.51	0.0655	0.9345	2.01	0.0222	0.9778	2.51	0.0060	0.9940
0.02	0.4920	0.5080	0.52	0.3015	0.6985	1.02	0.1539	0.8461	1.52	0.0643	0.9357	2.02	0.0217	0.9783	2.52	0.0059	0.9941
0.03	0.4880	0.5120	0.53	0.2981	0.7019	1.03	0.1515	0.8485	1.53	0.0630	0.9370	2.03	0.0212	0.9788	2.53	0.0057	0.9943
0.04	0.4840	0.5160	0.54	0.2946	0.7054	1.04	0.1492	0.8508	1.54	0.0618	0.9382	2.04	0.0207	0.9793	2.54	0.0055	0.9945
0.05	0.4801	0.5199	0.55	0.2912	0.7088	1.05	0.1469	0.8531	1.55	0.0606	0.9394	2.05	0.0202	0.9798	2.55	0.0054	0.9946
0.06	0.4761	0.5239	0.56	0.2877	0.7123	1.06	0.1446	0.8554	1.56	0.0594	0.9406	2.06	0.0197	0.9803	2.56	0.0052	0.9948
0.07	0.4721	0.5279	0.57	0.2843	0.7157	1.07	0.1423	0.8577	1.57	0.0582	0.9418	2.07	0.0192	0.9808	2.57	0.0051	0.9949
0.08	0.4681	0.5319	0.58	0.2810	0.7190	1.08	0.1401	0.8599	1.58	0.0571	0.9429	2.08	0.0188	0.9812	2.58	0.0049	0.9951
0.09	0.4641	0.5359	0.59	0.2776	0.7224	1.09	0.1379	0.8621	1.59	0.0559	0.9441	2.09	0.0183	0.9817	2.59	0.0048	0.9952
0.10	0.4602	0.5398	0.60	0.2743	0.7257	1.10	0.1357	0.8643	1.60	0.0548	0.9452	2.10	0.0179	0.9821	2.60	0.0047	0.9953
0.11	0.4562	0.5438	0.61	0.2709	0.7291	1.11	0.1335	0.8665	1.61	0.0537	0.9463	2.11	0.0174	0.9826	2.61	0.0045	0.9955
0.12	0.4522	0.5478	0.62	0.2676	0.7324	1.12	0.1314	0.8686	1.62	0.0526	0.9474	2.12	0.0170	0.9830	2.62	0.0044	0.9956
0.13	0.4483	0.5517	0.63	0.2643	0.7357	1.13	0.1292	0.8708	1.63	0.0516	0.9484	2.13	0.0166	0.9834	2.63	0.0043	0.9957
0.14	0.4443	0.5557	0.64	0.2611	0.7389	1.14	0.1271	0.8729	1.64	0.0505	0.9495	2.14	0.0162	0.9838	2.64	0.0041	0.9959
0.15	0.4404	0.5596	0.65	0.2578	0.7422	1.15	0.1251	0.8749	1.65	0.0495	0.9505	2.15	0.0158	0.9842	2.65	0.0040	0.9960
0.16	0.4364	0.5636	0.66	0.2546	0.7454	1.16	0.1230	0.8770	1.66	0.0485	0.9515	2.16	0.0154	0.9846	2.66	0.0039	0.9961
0.17	0.4325	0.5675	0.67	0.2514	0.7486	1.17	0.1210	0.8790	1.67	0.0475	0.9525	2.17	0.0150	0.9850	2.67	0.0038	0.9962
0.18	0.4286	0.5714	0.68	0.2483	0.7517	1.18	0.1190	0.8810	1.68	0.0465	0.9535	2.18	0.0146	0.9854	2.68	0.0037	0.9963
0.19	0.4247	0.5753	0.69	0.2451	0.7549	1.19	0.1170	0.8830	1.69	0.0455	0.9545	2.19	0.0143	0.9857	2.69	0.0036	0.9964
0.20	0.4207	0.5793	0.70	0.2420	0.7580	1.20	0.1151	0.8849	1.70	0.0446	0.9554	2.20	0.0139	0.9861	2.70	0.0035	0.9965
0.21	0.4168	0.5832	0.71	0.2389	0.7611	1.21	0.1131	0.8869	1.71	0.0436	0.9564	2.21	0.0136	0.9864	2.71	0.0034	0.9966
0.22	0.4129	0.5871	0.72	0.2358	0.7642	1.22	0.1112	0.8888	1.72	0.0427	0.9573	2.22	0.0132	0.9868	2.72	0.0033	0.9967
0.23	0.4090	0.5910	0.73	0.2327	0.7673	1.23	0.1093	0.8907	1.73	0.0418	0.9582	2.23	0.0129	0.9871	2.73	0.0032	0.9968
0.24	0.4052	0.5948	0.74	0.2296	0.7704	1.24	0.1075	0.8925	1.74	0.0409	0.9591	2.24	0.0125	0.9875	2.74	0.0031	0.9969
0.25	0.4013	0.5987	0.75	0.2266	0.7734	1.25	0.1056	0.8944	1.75	0.0401	0.9599	2.25	0.0122	0.9878	2.75	0.0030	0.9970
0.26	0.3974	0.6026	0.76	0.2236	0.7764	1.26	0.1038	0.8962	1.76	0.0392	0.9608	2.26	0.0119	0.9881	2.76	0.0029	0.9971
0.27	0.3936	0.6064	0.77	0.2206	0.7794	1.27	0.1020	0.8980	1.77	0.0384	0.9616	2.27	0.0116	0.9884	2.77	0.0028	0.9972
0.28	0.3897	0.6103	0.78	0.2177	0.7823	1.28	0.1003	0.8997	1.78	0.0375	0.9625	2.28	0.0113	0.9887	2.78	0.0027	0.9973
0.29	0.3859	0.6141	0.79	0.2148	0.7852	1.29	0.0985	0.9015	1.79	0.0367	0.9633	2.29	0.0110	0.9890	2.79	0.0026	0.9974
0.30	0.3821	0.6179	0.80	0.2119	0.7881	1.30	0.0968	0.9032	1.80	0.0359	0.9641	2.30	0.0107	0.9893	2.80	0.0026	0.9974
0.31	0.3783	0.6217	0.81	0.2090	0.7910	1.31	0.0951	0.9049	1.81	0.0351	0.9649	2.31	0.0104	0.9896	2.81	0.0025	0.9975
0.32	0.3745	0.6255	0.82	0.2061	0.7939	1.32	0.0934	0.9066	1.82	0.0344	0.9656	2.32	0.0102	0.9898	2.82	0.0024	0.9976
0.33	0.3707	0.6293	0.83	0.2033	0.7967	1.33	0.0918	0.9082	1.83	0.0336	0.9664	2.33	0.0099	0.9901	2.83	0.0023	0.9977
0.34	0.3669	0.6331	0.84	0.2005	0.7995	1.34	0.0901	0.9099	1.84	0.0329	0.9671	2.34	0.0096	0.9904	2.84	0.0023	0.9977
0.35	0.3632	0.6368	0.85	0.1977	0.8023	1.35	0.0885	0.9115	1.85	0.0322	0.9678	2.35	0.0094	0.9906	2.85	0.0022	0.9978
0.36	0.3594	0.6406	0.86	0.1949	0.8051	1.36	0.0869	0.9131	1.86	0.0314	0.9686	2.36	0.0091	0.9909	2.86	0.0021	0.9979
0.37	0.3557	0.6443	0.87	0.1922	0.8078	1.37	0.0853	0.9147	1.87	0.0307	0.9693	2.37	0.0089	0.9911	2.87	0.0021	0.9979
0.38	0.3520	0.6480	0.88	0.1894	0.8106	1.38	0.0838	0.9162	1.88	0.0301	0.9699	2.38	0.0087	0.9913	2.88	0.0020	0.9980
0.39	0.3483	0.6517	0.89	0.1867	0.8133	1.39	0.0823	0.9177	1.89	0.0294	0.9706	2.39	0.0084	0.9916	2.89	0.0019	0.9981
0.40	0.3446	0.6554	0.90	0.1841	0.8159	1.40	0.0808	0.9192	1.90	0.0287	0.9713	2.40	0.0082	0.9918	2.90	0.0019	0.9981
0.41	0.3409	0.6591	0.91	0.1814	0.8186	1.41	0.0793	0.9207	1.91	0.0281	0.9719	2.41	0.0080	0.9920	2.91	0.0018	0.9982
0.42	0.3372	0.6628	0.92	0.1788	0.8212	1.42	0.0778	0.9222	1.92	0.0274	0.9726	2.42	0.0078	0.9922	2.92	0.0018	0.9982
0.43	0.3336	0.6664	0.93	0.1762	0.8238	1.43	0.0764	0.9236	1.93	0.0268	0.9732	2.43	0.0075	0.9925	2.93	0.0017	0.9983
0.44	0.3300	0.6700	0.94	0.1736	0.8264	1.44	0.0749	0.9251	1.94	0.0262	0.9738	2.44	0.0073	0.9927	2.94	0.0016	0.9984
0.45	0.3264	0.6736	0.95	0.1711	0.8289	1.45	0.0735	0.9265	1.95	0.0256	0.9744	2.45	0.0071	0.9929	2.95	0.0016	0.9984
0.46	0.3228	0.6772	0.96	0.1685	0.8315	1.46	0.0721	0.9279	1.96	0.0250	0.9750	2.46	0.0069	0.9931	2.96	0.0015	0.9985
0.47	0.3192	0.6808	0.97	0.1660	0.8340	1.47	0.0708	0.9292	1.97	0.0244	0.9756	2.47	0.0068	0.9932	2.97	0.0015	0.9985
0.48	0.3156	0.6844	0.98	0.1635	0.8365	1.48	0.0694	0.9306	1.98	0.0239	0.9761	2.48	0.0066	0.9934	2.98	0.0014	0.9986
0.49	0.3121	0.6879	0.99	0.1611	0.8389	1.49	0.0681	0.9319	1.99	0.0233	0.9767	2.49	0.0064	0.9936	2.99	0.0014	0.9986
0.50	0.3085	0.6915	1.00	0.1587	0.8413	1.50	0.0668	0.9332	2.00	0.0228	0.9772	2.50	0.0062	0.9938	3.00	0.0013	0.9987

65

Exemplo

Retornem-se os dados do exemplo sobre as cores favoritas das pessoas (do teste de Qui-quadrado de independência). Identificar quais são as cores preferida que mais contribuem para a dependência.

Resolução

Frequências observadas (O_{ij})

	Pre.	Bra.	Ver.	Az.	Tot.
M	48	12	33	57	**150**
H	35	46	42	27	**150**
Tot.	**83**	**58**	**75**	**84**	**300**

Frequências esperadas (E_{ij})

	Pre.	Bra.	Ver.	Az.	Tot.
M	41,5	29	37,5	42	**150**
H	41,5	29	37,5	42	**150**
Tot.	**83**	**58**	**75**	**84**	**300**

Calculo dos resíduos

$$R_{11} = \frac{O_{11} - E_{11}}{\sqrt{E_{11}\left(1 - \frac{O_{1.}}{n}\right)\left(1 - \frac{O_{.1}}{n}\right)}} = \frac{48 - 41{,}5}{\sqrt{41{,}5\left(1 - \frac{150}{300}\right)\left(1 - \frac{83}{300}\right)}} \approx 1{,}68$$

$$\vdots$$

$$R_{24} = \frac{O_{24} - E_{24}}{\sqrt{E_{24}\left(1 - \frac{O_{2.}}{n}\right)\left(1 - \frac{O_{.4}}{n}\right)}} = \frac{27 - 42}{\sqrt{42\left(1 - \frac{150}{300}\right)\left(1 - \frac{84}{300}\right)}} \approx -3{,}86$$

66

Resíduos (R_{ij})

	Pre.	Bra.	Ver.	Az.
M	1,68	-4,97	-1,20	3,86
H	-1,68	4,97	1,20	-3,86

Determinar o valor de $Z_{1-\frac{\alpha}{2}}$

$\alpha = 0{,}05 \implies Z_{1-\frac{\alpha}{2}} = 1{,}96$

Vimos que $\{R_{12}, R_{22}, R_{14}, R_{24}\} > Z_{1-\frac{\alpha}{2}}$, logo as células que contribuem para dependência são: (1,2), (2,2), (1,4) e (2,4).

Como R_{12} e R_{22} têm maior valor absoluto (i.e., 4,97), então as células que mais contribuem para a dependência das variáveis são as células (1,2) e (2,2), isto é, a cor branca é que mais contribui para a dependência.

67

3.2. Modelo log-lineares

A análise log-linear de tabela contingência permite:

- Averiguar a existência (ou não) de dependência entre as variáveis,
- Quantificar os efeitos que as variáveis ou a sua combinação exercem sobre os resultados observados.

Objetivo: ajustar modelos que caracterizam, tão bem quanto possível, a estrutura subjacente dos dados.

68

a) Modelo independência

Supondo H_0 de independência verdadeira,

$$E_{ij} = \frac{O_{i.} \times O_{.j}}{n}, i = \{1, \dots, r\} \text{ e j} = \{1, \dots, c\}$$

O modelo de logaritmo é dado por:

$$\ln E_{ij} = \ln\left(\frac{O_{i.} \times O_{j.}}{n}\right) = \ln O_{i.} + \ln O_{.j} - \ln n$$

Isto é, se as variáveis forem independentes, o logaritmo natural da frequência esperada E_{ij} é a soma de

- efeito linha i (variável A),
- efeito coluna j (variável B),
- efeito constante.

69

Assim, o modelo log-linear de independência deve ser escrito da seguinte forma:

$$\ln E_{ij} = \mu + \lambda_i^A + \lambda_j^B \quad \text{..... (*)}$$

Para estimar os parâmetros do modelo é necessário considerar

$$\sum_{i=1}^{r} \lambda_i^A = 0 \quad e \quad \sum_{j=1}^{c} \lambda_j^B = 0,$$

e assim os parâmetros do modelo são estimados através de

- $\hat{\mu} = \frac{1}{r \times c} \sum_{i=1}^{r} \sum_{j=1}^{c} \ln E_{ij}$; efeito médio global
- $\hat{\lambda}_i^A = \frac{1}{c} \sum_{j=1}^{c} \ln E_{ij} - \mu$; efeito principal da variável A
- $\hat{\lambda}_j^B = \frac{1}{r} \sum_{i=1}^{r} \ln E_{ij} - \mu$; efeito principal da variável B

Como o número de parâmetros é respetivamente 1, $r-1$ e $c-1$, logo, o número total dos parâmetros independentes é $r + c - 1$.

70

b) Modelo saturado (modelo completo)

Se não houver independência, o modelo anterior (*) torna-se inadequado, sendo necessário introduzir um termo representativo da interação entre as variáveis, seja λ_{ij}^{AB}, e o modelo fica

$$\ln E_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \quad \text{..... (**)}$$

Neste caso, $E_{ij} \neq \frac{O_i \times O_j}{n}$, pois H_0 de independência não é verdadeira

$$E_{ij} = n \times P(A_i \cap B_j) = n \times \frac{O_{ij}}{n} = O_{ij}$$

Com as restrições

$$\sum_{i=1}^{r} \lambda_i^A = 0, \sum_{j=1}^{c} \lambda_j^B = 0 \text{ e } \sum_{i=1}^{r} \lambda_{ij}^{AB} = \sum_{j=1}^{c} \lambda_{ij}^{AB} = 0$$

os parâmetros do modelo são estimados de forma equivalente aos parâmetros do modelo (*) substituindo E_{ij} por O_{ij} e adicionalmente $\hat{\lambda}_{ij}^{AB} = \ln O_{ij} - (\mu + \lambda_i^A + \lambda_j^B)$.

O número total de parâmetros independentes $1 + (r-1) + (c-1) + (r-1) \times (c-1) = r \times c$

71

Tipos de modelos

- *Modelo abrangentes*: inclui pelo menos todos os parâmetros relativos aos efeitos principais de cada uma das variáveis, isto é:
  - $\ln E_{ij} = \mu + \lambda_i^A + \lambda_j^B$
  - $\ln E_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$

- *Modelo não-abrangentes* (*noncomprehensive*): não inclui pelo menos um parâmetro do efeito principal
  - Se $\lambda_j^B = 0 \Longrightarrow \ln E_{ij} = \mu + \lambda_i^A$, $\forall i, j \longrightarrow$ efeito de variável B é nulo, isto é, as c categorias da variável B são equiprováveis (têm a mesma probabilidade).
  - Se $\lambda_i^A = 0 \Longrightarrow \ln E_{ij} = \mu + \lambda_j^B$, $\forall i, j \longrightarrow$ efeito de variável A é nulo
  - Se $\lambda_i^A = \lambda_j^B = 0 \Longrightarrow \ln E_{ij} = \mu$, $\forall i, j \longrightarrow$ só existe um efeito constante, que não depende nem de A nem de B.

72

Interpretação dos parâmetros do modelo

- Se $\lambda_i^A > 0\ (< 0)$, o efeito da linha-*i* é positivo (negativo), i.e. as frequências esperadas da linha-*i* tendem a ser superiores (inferiores) à média global (μ).

- Se $\lambda_j^B > 0\ (< 0)$, o efeito da coluna-*j* é positivo (negativo), i.e. as frequências esperadas da coluna-*j* tendem a ser superiores (inferiores) à média global (μ).

- Se $\lambda_{ij}^{AB} > 0\ (< 0)$, há uma associação positiva (negativa) entre A e B. Para $\lambda_{ij}^{AB} = 0$, não existe associação entre A e B.

73

Por exemplo, considere-se o caso em que, $\lambda_i^A > 0$.

$$\lambda_i^A > 0 \Leftrightarrow \frac{1}{c}\sum_{j=1}^{c} \ln E_{ij} - \mu > 0 \Leftrightarrow \frac{1}{c}\sum_{j=1}^{c} \ln E_{ij} > \mu$$

Assim, $\lambda_i^A > 0$ indica que o efeito da linha-i é positiva, isto é a média do ln das frequências esperadas é superior à média global μ.

Ou de outra forma,

$$\frac{1}{c}\sum_{j=1}^{c} \ln E_{ij} - \mu = \frac{1}{c}\sum_{j=1}^{c} \ln E_{ij} - \frac{c}{c}\mu = \frac{1}{c}\left(\sum_{j=1}^{c} \ln E_{ij} - c\mu\right) = \frac{1}{c}\left(\sum_{j=1}^{c} \ln E_{ij} - \sum_{j=1}^{c} \mu\right) = \frac{1}{c}\left(\sum_{j=1}^{c} (\ln E_{ij} - \mu)\right)$$

Assim, para $\lambda_i^A > 0$,

$$\frac{1}{c}\left(\sum_{j=1}^{c} (\ln E_{ij} - \mu)\right) > 0$$

isto é, o ln das frequências esperadas é (em média) superior à média global μ.

74

Ajustamento de modelos log-lineares

Objetivo: verificar qual dos modelos, o de independência ou o saturado, melhor se ajusta aos dados.

H_0: modelo de independência ($\lambda_{ij}^{AB} = 0$)
H_1: modelo saturado ($\lambda_{ij}^{AB} \neq 0$)

Passos para ajustar os modelos log-lineares:

- Aplicar o teste do Qui-quadrado (χ^2) ou o teste de razão de verosimilhança (G^2), supondo H_0 verdadeiro. Se H_0 for rejeitada, considera-se o modelo saturado. Caso contrário, considera-se o modelo de independência.
- Estimar os parâmetros do modelo a considerar.
- Interpretar os resultados obtidos.

75

Exemplo 1

Retornem-se os dados do exemplo sobre as cores favoritas das pessoas (do teste de Qui-quadrado de independência), onde se concluir a existência de forte associação. Assim, o modelo adequado é o ***saturado***.

Estimativas dos parâmetros

Célula (i,j)	$ln\ E_{ij}$	μ	λ_i^A		λ_j^B				λ_{ij}^{AB}
			λ_1^A	λ_2^A	λ_1^B	λ_2^B	λ_3^B	λ_4^B	
(1,1)	3,87	3,54	-0,07	0,07	0,17	-0,38	0,08	0,13	0,22
(1,2)	2,48								-0,61
(1,3)	3,50								-0,06
(1,4)	4,04								0,44
(2,1)	3,56								-0,22
(2,2)	3,83								0,61
(2,3)	3,74								0,06
(2,4)	3,30								-0,44
Total	**28,31**								

O modelo:

$$\ln E_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

Por exemplo,

$\ln E_{12} = \mu + \lambda_1^A + \lambda_2^B + \lambda_{12}^{AB}$

$2{,}48 = 3{,}54 - 0{,}07 - 0{,}38 - 0{,}61$

$2{,}48 = 2{,}48$ (verdadeiro)

76

Interpretação dos resultados:

- $\mu = 3{,}54 \longrightarrow$ o efeito médio global é positivo.
- Os valor dos efeitos de género $\boldsymbol{\lambda_1^A}$ (mulher) e $\boldsymbol{\lambda_2^A}$ (homem) são simétricos (esta é a uma relação esperada uma vez que existem apenas duas classes em A, e estas têm igual probabilidade).
- Os efeitos das cores favoritas $\boldsymbol{\lambda_1^B}$ (preta), $\boldsymbol{\lambda_3^B}$ (vermelha) e $\boldsymbol{\lambda_4^B}$ (azul) são positivos com maior efeito para a cor azul. Por outro lado, $\boldsymbol{\lambda_2^B}$ (branca) tem efeito negativo.
- Os efeitos de intersecção são simétricos 2 a 2, isto é: $\boldsymbol{\lambda_{11}^{AB}}$ e $\boldsymbol{\lambda_{21}^{AB}}$, o efeito mulher/homem que têm cor favorita preta é simétrico. Assim como, os pares $\boldsymbol{\lambda_{12}^{AB}}$ e $\boldsymbol{\lambda_{22}^{AB}}$; $\boldsymbol{\lambda_{13}^{AB}}$ e $\boldsymbol{\lambda_{23}^{AB}}$; $\boldsymbol{\lambda_{14}^{AB}}$ e $\boldsymbol{\lambda_{24}^{AB}}$.
- Nota-se $\boldsymbol{\lambda_{12}^{AB}}$ tem o maior peso negativo e $\boldsymbol{\lambda_{22}^{AB}}$ tem maior peso positivo.

77 **Exemplo 2**

A tabela seguinte mostra os resultados de uma avaliação de satisfação com a compra de um novo modelo de automóvel de luxo.

a) Teste a hipótese de que o resultado de avaliação depende do género.

b) Determine o modelo log-linear que se ajusta aos dados.

Consumidores	Avaliação de satisfação			Total
	Muito	Pouco	Não	
Homem	30	15	15	**60**
Mulher	25	10	5	**40**
Total	**55**	**25**	**20**	**100**

78 **Resolução do exemplo 2**

a) Teste de independência

- Hipóteses a testar:

  H_0: o resultado de avaliação não depende do género vs. H_1: $\sim H_0$

Frequências observadas

Cons.	Avaliação de satisfação			Total
	Muito	Pouco	Não	
Homem	30	15	15	**60**
Mulher	25	10	5	**40**
Total	**55**	**25**	**20**	**100**

Frequências esperadas

Cons.	Avaliação de satisfação			Total
	Muito	Pouco	Não	
Homem	33	15	12	**60**
Mulher	22	10	8	**40**
Total	**55**	**25**	**20**	**100**

- Estatística do teste

  $\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(30-33)^2}{33} + \cdots + \frac{(5-8)^2}{8} \approx 2{,}56$

- Valor crítico: $\alpha = 0{,}05, r = 2, c = 3 \Rightarrow \chi^2_{(1-\alpha);(r-1)\times(c-1)} = \chi^2_{0{,}95;2} \approx 5{,}99$

- Decisão: Como $\boldsymbol{\chi^2_{obs} < \chi^2_{0{,}95;2}}$, não se rejeita H_0. Conclui-se, ao nível significância 5%, que o resultado de avaliação de satisfação não depende do género.

79

b) Como se concluiu não rejeitar H_0, considera-se o modelo de independência.

- Estimação dos parâmetros do modelo

Célula (i,j)	$ln\ E_{ij}$	μ	λ_i^A		λ_j^B			λ_{ij}^{AB}
			λ_1^A	λ_2^A	λ_1^B	λ_2^B	λ_3^B	
(1,1)	3,50	2,69	0,20	-0,20	0,60	-0,19	-0,41	0,00
(2,1)	3,09							0,00
(1,2)	2,71							0,00
(2,2)	2,30							0,00
(1,3)	2,48							0,00
(2,3)	2,08							0,00
total	**16,16**							

O modelo:

$$\ln E_{ij} = \mu + \lambda_i^A + \lambda_j^B$$

Por exemplo,

$\ln E_{12} = \mu + \lambda_1^A + \lambda_2^B$

$2{,}71 = 2{,}69 + 0{,}20 - 0{,}19$

$2{,}71 = 2{,}71$ (verdadeiro)

80

- Interpretação dos resultados:

- $\mu = 2{,}69 \longrightarrow$ o efeito médio global é positivo.
- Os valor dos efeitos de género $\boldsymbol{\lambda_1^A}$ (homem) e $\boldsymbol{\lambda_2^A}$ (mulher) são simétricos.
- Os efeitos dos resultados de avaliação $\boldsymbol{\lambda_1^B}$ (muito satisfaz) é positivo com maior peso.
- $\boldsymbol{\lambda_2^B}$ (pouco satisfaz) e $\boldsymbol{\lambda_3^B}$ (não satisfaz) têm efeitos negativos.
- Os efeitos de intersecção são nulos, porque as variáveis são independentes.

81

4. Tabelas de contingência $r \times c \times l$

Se uma tabela tem r linhas, c colunas e l estratos diz-se que tem dimensão $r \times c \times l$, e designa-se por tabela *tridimensional*.

	C_1	...	C_l	Tot.
	B_1 ... B_c	...	B_1 ... B_c	
A_1	O_{111} ... O_{1c1}	...	O_{11l} ... O_{1cl}	$O_{i..}$
.	. .		. .	.
.	. .		. .	.
.	. .		. .	.
A_r	O_{r11} ... O_{rc1}	...	O_{r1l} ... O_{rcl}	$O_{r..}$
Tot.	$O_{.1.}$ ... $O_{.c.}$	...	$O_{.1.}$ ... $O_{.c.}$	n

- Dimensão da amostra:

$$n = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} O_{ijk}$$

- Total marginal de uma só variável:

$$O_{i..} = \sum_{j=1}^{c}\sum_{k=1}^{l} O_{ijk}\,; i = 1, \dots, r \qquad O_{.j.} = \sum_{i=1}^{r}\sum_{k=1}^{l} O_{ijk}\,; j = 1, \dots, c$$

$$O_{..k} = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ijk}\,; k = 1, \dots, l$$

- Total marginal de duas variáveis:

$$O_{ij.} = \sum_{k=1}^{l} O_{ijk}\,; i = 1, \dots, r; \mathrm{j} = 1, \dots, c \qquad O_{i.k} = \sum_{j=1}^{c} O_{ijk}\,; i = 1, \dots, c; k = 1, \dots, l \qquad O_{.jk} = \sum_{i=1}^{r} O_{ijk}\,; j = 1, \dots, r; k = 1, \dots, l$$

82

4.1. Hipóteses de Independência

Em tabelas tridimensionais teremos de testar mais do que uma hipótese nula,

- Independência mútua ⟶ Três variáveis são independentes, ou não estão associados entre si.
- Independência parcial ⟶ Duas variáveis são independentes relativamente à terceira.
- Independência condicional ⟶ Duas variáveis são independentes para uma categoria específica da terceira variável, i.e., a terceira variável é controlada.
- Associação 2 a 2 ⟶ A relação parcial condicional entre quaisquer duas variáveis é a mesma para cada nível da terceira variável.

83

a) Independência Mútua

Se existe independência mútua entre A, B e C então, $P(A \cap B \cap C)) = P(A) \times P(B) \times P(C)$

H_0: as três variáveis são independentes

H_1: $\sim H_0$

Supondo H_0 de independência mútua verdadeira,

$$E_{ijk} = n \times P(A_i \cap B_j \cap C_k) = n \times P(A_i) \times P(B_j) \times P(C_k) = n \times \frac{O_{i..}}{n} \times \frac{O_{.j.}}{n} \times \frac{O_{..k}}{n} = \frac{O_{i..} \times O_{.j.} \times O_{..k}}{n^2}, \forall ijk$$

84

b) Independência parcial

Se existe independência parcial entre C e AB então, $P(C \cap (A \cap B)) = P(C) \times P(A \cap B)$

Existem 3 hipótese para testar independência parcial.

– Independência parcial entre C e (AB)

$H_0^{(1)}$: a variável C independente das restantes (AB)

$H_1^{(1)}$: $\sim H_0^{(1)}$

Supondo $H_0^{(1)}$ verdadeira,

$$E_{ijk} = n \times P(A_i \cap B_j \cap C_k) = n \times P((A_i \cap B_j) \cap C_k) = n \times P(A_i \cap B_j) \times P(C_k) = n \times \frac{O_{ij.}}{n} \times \frac{O_{..k}}{n} = \frac{O_{ij.} \times O_{..k}}{n}; \forall i, j, k$$

85

– Independência parcial entre B e (AC)

$H_0^{(2)}$: a variável B independente das restantes (AC)

$H_1^{(2)}: \sim H_0^{(2)}$

Supondo $H_0^{(2)}$ verdadeira, $E_{ijk} = \frac{O_{i.k} \times O_{.j.}}{n}; \forall i, j, k$

– Independência parcial entre A e (BC)

$H_0^{(3)}$: a variável A independente das restantes (BC)

$H_1^{(3)}: \sim H_0^{(3)}$

Supondo $H_0^{(3)}$ verdadeira, $E_{ijk} = \frac{O_{.jk} \times O_{i..}}{n}; \forall i, j, k$

86

c) Independência condicional

Existem 3 hipótese para testar independência condicional.

Independência condicional,

- $P(A|C) = P(A)$
- $P(A \cap B|C) = P(A|C) \times P(B|C)$

– Independência condicional entre A e B dada C

$H_0^{(1)}$: as variáveis A e B são condicionalmente independentes de C

$H_1^{(1)}: \sim H_0^{(1)}$

Supondo $H_0^{(1)}$ verdadeira,

$$E_{ijk} = n \times P(A_i \cap B_j \cap C_k) = n \times P(A_i \cap B_j | C_k) \times P(C_k) = n \times P(A_i|C_k) \times P(B_j|C_k) \times P(C_k) =$$

$$= n \times \frac{P(A_i \cap C_k)}{P(C_k)} \times \frac{P(B_j \cap C_k)}{P(C_k)} \times P(C_k) = n \times \frac{\frac{O_{i.k}}{n}}{\frac{O_{..k}}{n}} \times \frac{O_{.jk}}{n} = \frac{O_{i.k} \times O_{.jk}}{O_{..k}}; \forall i, j, k$$

87

– Independência condicional entre A e C dada B

$H_0^{(2)}$: as variáveis A e C são independente dada variável B

$H_1^{(2)}: \sim H_0^{(2)}$

Supondo $H_0^{(2)}$ verdadeira, $E_{ijk} = \frac{O_{ij.} \times O_{.jk}}{O_{.j.}}; \forall i, j, k$

– Independência condicional entre B e C dada A

$H_0^{(3)}$: as variáveis B e C são independente dada variável A

$H_1^{(3)}: \sim H_0^{(3)}$

Supondo $H_0^{(3)}$ verdadeira, $E_{ijk} = \frac{O_{ij.} \times O_{i.k}}{O_{i..}}; \forall i, j, k$

88

d) Associação 2 a 2

Hipóteses a testar

H_0: a associação entre duas variáveis não depende das categorias da terceira

$H_1: \sim H_0$

Não é possível escrever explicitamente E_{ijk} em função dos valores de $O_{...}$ e é necessário utilizar um método iterativo para estimar E_{ijk}.

89

Método iterativo: ajustamento proporcional iterativo

Este método permite obter as estimativas das frequências esperadas E_{ijk} em modelo log-lineares hierárquicos. Este método iterativo começa por usar quaisquer estimativas iniciais, $\left\{\hat{E}_{ijk}^{(0)}\right\}$, desde que satisfaçam o modelo ajustar. Multiplicando estes valores por fatores de escala apropriados, ajustam-se sucessivamente as estimativas iniciais de modo a que os seus valores coincidam com as frequências marginais que consistem um conjunto mínimo de informação que é "suficiente" para obter E_{ijk} no modelo.

90

Passos do método de ajustamento proporcional iterativo

Considere-se a tabela inicial O_{ijk} e os totais marginais

Passo 0: estimar as frequências esperadas iniciais $\hat{E}_{ijk}^{(0)}$ e calcular os totais marginais

Passo 1: (neste caso para o modelo (AB,AC,BC))

- Ajustar linha/coluna
  - $\hat{E}'_{ijk} = \frac{O_{ijk} \times \hat{E}_{ij.}^{(0)}}{O_{ij.}}$
  - $\hat{E}_{ijk}^{(iv)} = \frac{\hat{E}'''_{ijk} \times \hat{E}_{ij.}^{(0)}}{\hat{E}'''_{ij.}}$
  - .
  - .
  - .
- Ajustar linha/estratos
  - $\hat{E}''_{ijk} = \frac{\hat{E}'_{ijk} \times \hat{E}_{i.k}^{(0)}}{\hat{E}'_{i.k}}$
  - $\hat{E}_{ijk}^{(v)} = \frac{\hat{E}_{ijk}^{(iv)} \times \hat{E}_{i.k}^{(0)}}{\hat{E}_{i.k}^{(iv)}}$
  - .
  - .
  - .
- Ajustar coluna/estrato
  - $\hat{E}'''_{ijk} = \frac{\hat{E}''_{ijk} \times \hat{E}_{.jk}^{(0)}}{\hat{E}''_{.jk}}$
  - $\hat{E}_{ijk}^{(vi)} = \frac{\hat{E}_{ijk}^{(v)} \times \hat{E}_{.jk}^{(0)}}{\hat{E}_{.jk}^{(v)}}$
  - .
  - .
  - .

E assim sucessivamente até os totais marginais das linhas e colunas coincidirem com os totais marginais obtidas da estimação inicial das frequências esperadas $\hat{E}_{ijk}^{(0)}$.

91

Exemplo de aplicação do método iterativo (usando Excel)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1																			
2			Frequências observadas (Oijk)								Frequências esperadas iniciais (Eijk)								
3			Estratos								Estratos								
4	Linha	Coluna	1º	2º	3º	4º			Linha	Coluna	1º	2º	3º	4º					
5	1ª	1ª	5	4	6	5	20		1ª	1ª	5	7	4	5	21				
6	1ª	2ª	3	6	5	2	16		1ª	2ª	4	6	5	2	17				
7	2ª	1ª	8	4	3	1	16		2ª	1ª	8	4	7	3	22				
8	2ª	2ª	2	5	1	3	11		2ª	2ª	4	3	4	5	16				
9	3ª	1ª	6	4	3	6	19		3ª	1ª	6	4	3	4	17				
10	3ª	2ª	2	8	2	6	18		3ª	2ª	5	8	4	6	23				
11			26	31	20	23	100				32	32	27	25	116				
12																			
13	1		8	10	11	7	36		1		9	13	9	7	38				
14	2		10	9	4	4	27		2		12	7	11	8	38				
15	3		8	12	5	12	37		3		11	12	7	10	40				
16																			
17		1	19	12	12	12	55			1	19	15	14	12	60				
18		2	7	19	8	11	45			2	13	17	13	13	56				

Frequências observadas/esperadas

Totais marginais de duas variáveis

Totais marginais de uma só variável

Total de observações/esperadas

Totais marginais:

- Uma só variável:
  - $O_{i..}$(G13:G15), por exemplo, G13=soma(C13:F13)
  - $O_{.j.}$(G17:G18), por exemplo, G17=soma(C17:F17)
  - $O_{..k}$(C11:F11), por exemplo, C11=soma(C5:C10)
- Duas variáveis:
  - $O_{ij.}$(G5:G10), por exemplo, G5=soma(C5:F5)
  - $O_{i.k}$(C13:F15), por exemplo, C13=soma(C5;C6)
  - $O_{.jk}$(C17:F18), por exemplo, C17=soma(C5;C7;C9)

92

1ª iteração

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
21			Ajustar linha/coluna							Ajustar linha/estratos						Ajustar coluna/estratos			
22																			
23	1	1	5,25	4,20	6,30	5,25	21,00		5,60	5,16	4,88	4,98	20,63		4,72	6,86	4,12	5,43	21,12
24	1	2	3,19	6,38	5,31	2,13	17,00		3,40	7,84	4,12	2,02	17,37		4,67	6,43	5,15	1,88	18,13
25	2	1	11,00	5,50	4,13	1,38	22,00		9,49	3,01	8,13	1,92	22,55		8,00	4,01	6,86	2,09	20,95
26	2	2	2,91	7,27	1,45	4,36	16,00		2,51	3,99	2,87	6,08	15,45		3,45	3,27	3,58	5,66	15,96
27	3	1	5,37	3,58	2,68	5,37	17,00		7,45	3,11	3,59	4,12	18,27		6,28	4,13	3,02	4,49	17,93
28	3	2	2,56	10,22	2,56	7,67	23,00		3,55	8,89	3,41	5,88	21,73		4,88	7,30	4,27	5,47	21,91
29			30,27	37,15	22,43	26,15	116,00		32,00	32,00	27,00	25,00	116,00		32,00	32,00	27,00	25,00	116,00
30																			
31	1		8,44	10,58	11,61	7,38	38,00		9,00	13,00	9,00	7,00	38,00		9,39	13,29	9,26	7,30	39,25
32	2		13,91	12,77	5,58	5,74	38,00		12,00	7,00	11,00	8,00	38,00		11,45	7,28	10,44	7,74	36,91
33	3		7,92	13,80	5,24	13,04	40,00		11,00	12,00	7,00	10,00	40,00		11,16	11,43	7,29	9,95	39,83
34																			
35		1	21,62	13,28	13,11	11,99	60,00		22,54	11,29	16,60	11,02	61,45		19,00	15,00	14,00	12,00	60,00
36		2	8,65	23,87	9,32	14,16	56,00		9,46	20,71	10,40	13,98	54,55		13,00	17,00	13,00	13,00	56,00

C23=C5*O5/G5

I23=C23*K13/C23

O23=I23*K17/I35

Repetir o processo até obter os totais marginais semelhantes aos totais marginais das frequências esperadas iniciais $\hat{E}_{ijk}^{(0)}$.

...

93 4ª iteração

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
68			Ajustar linha/coluna							Ajustar linha/estratos						Ajustar coluna/estratos			
69																			
70	1	1	4,64	6,95	4,07	5,34	21,00		4,64	6,95	4,07	5,34	21,00		4,64	6,95	4,07	5,34	21,00
71	1	2	4,36	6,05	4,93	1,66	17,00		4,36	6,05	4,93	1,66	17,00		4,36	6,05	4,93	1,66	17,00
72	2	1	8,48	3,96	7,25	2,30	22,00		8,48	3,96	7,26	2,30	22,00		8,48	3,96	7,26	2,30	22,00
73	2	2	3,52	3,05	3,73	5,70	16,00		3,52	3,04	3,74	5,70	16,00		3,52	3,04	3,74	5,70	16,00
74	3	1	5,88	4,08	2,68	4,35	17,00		5,88	4,09	2,67	4,35	17,00		5,88	4,09	2,67	4,36	17,00
75	3	2	5,12	7,90	4,33	5,65	23,00		5,12	7,91	4,33	5,65	23,00		5,12	7,91	4,33	5,64	23,00
76			32,00	32,00	27,00	25,00	116,00		32,00	32,00	27,00	25,00	116,00		32,00	32,00	27,00	25,00	116,00
77																			
78	1		8,99	13,00	9,00	7,01	38,00		9,00	13,00	9,00	7,00	38,00		9,00	13,00	9,00	7,00	38,00
79	2		12,00	7,01	10,99	8,00	38,00		12,00	7,00	11,00	8,00	38,00		12,00	7,00	11,00	8,00	38,00
80	3		11,00	11,98	7,01	10,00	40,00		11,00	12,00	7,00	10,00	40,00		11,00	12,00	7,00	10,00	40,00
81																			
82		1	19,00	15,00	14,00	12,00	60,00		19,00	15,00	14,00	11,99	60,00		19,00	15,00	14,00	12,00	60,00
83		2	13,00	17,00	13,00	13,00	56,00		13,00	17,00	13,00	13,01	56,00		13,00	17,00	13,00	13,00	56,00

C70=O53*O5/S53 I70=C70*K13/C78 O70=I70*K17/I82

O processo termina na 4ª iteração, na qual os totais marginais na última tabela (i.e., tabela de ajustar coluna/estratos) são iguais aos totais marginais das frequências esperadas iniciais $\hat{E}_{ijk}^{(0)}$.

Assim, os valores das células na última tabela (*i.e. O70:R75*) são considerados como os valores das frequências esperadas do modelo (AB,AC,BC).

94

A estatística de teste: $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\frac{(O_{ijk}-E_{ijk})^2}{E_{ijk}}$

Quando o tamanho de amostra é elevado, χ^2 segue aproximadamente uma distribuição Qui-quadrado com graus de liberdade :

- Independência mútua: $rcl - r - c - l + 2$
- Independência parcial entre (AB) e C: $(rc-1)(l-1)$
- Independência parcial entre (AC) e B: $(rl-1)(c-1)$
- Independência parcial entre (BC) e A: $(cl-1)(r-1)$
- Independência condicional entre A e B dada C: $l(r-1)(c-1)$
- Independência condicional entre A e C dada B: $c(r-1)(l-1)$
- Independência condicional entre B e C dada A: $r(c-1)(l-1)$
- Associação 2 a 2: $(r-1)(l-1)(c-1)$

Decisão: Rejeita-se H_0 (em favor de H_1) quando χ^2_{obs} é maior ou igual ao valor crítico do teste, caso contrário não se rejeita H_0 em favor de H_1.

95

Os passos do teste χ^2 para tabela tridimensional:

1) Determinar as hipóteses a testar
   H_0: as variáveis não estão associados (i.e. são independentes)
   H_1: $\sim H_0$
2) Calcular os totais marginais
3) Estimar as frequências esperadas E_{ijk} do modelo, supondo H_0 verdadeira
4) Calcular o valor observado da estatística do teste χ^2
5) Calcular os graus de liberdade (gl) do modelo
6) Determinar o ponto crítico
7) Decisão: Rejeita-se H_0 (em favor de H_1) quando $\chi^2_{obs} \geq \boldsymbol{\chi^2_{(1-\alpha);gl}}$. Caso contrário, não se rejeita H_0.

96

Exemplo

A tabela seguinte apresenta o número de respondentes classificados de acordo com o género, o partido político e a ideologia política.

Género	Partido político	Ideologia política				
		Muito liberal	Levemente liberal	Moderada	Levemente conservadora	Muito conservadora
M	Democrata	44	47	118	23	32
	Republicano	18	28	86	39	48
H	Democrata	36	34	53	18	23
	Republicano	12	18	62	45	51

Testar se o género, o partido político e a ideologia política estão associados.

97

Resolução do exemplo

Gén.	P.P	Ideologia política				
		ML	LL	Mod	LC	MC
M	D	44	47	118	23	32
	R	18	28	86	39	48
H	D	36	34	53	18	23
	R	12	18	62	45	51

❑ *Hipóteses a testar:*

H_0: o género, o partido político e a ideologia política não estão associados (ou são independentes)

$H_1: \sim H_0$

❑ *Os totais marginais*

- *De uma só variável:*

Género	
M	483
H	352

Part. pol.	
D	428
R	407

Ideologia política				
ML	LL	Mod	LC	MC
110	127	319	125	154

- *De duas variáveis:*

Gén.	P.P.	
	D	R
M	264	219
H	164	188

Gén	Ideologia política				
	ML	LL	Mod	LC	MC
M	62	75	204	62	80
H	48	52	115	63	74

P.P.	Ideologia política				
	ML	LL	Mod	LC	MC
D	80	81	171	41	55
R	30	46	148	84	99

98

❑ *Total de observação:* $n = 835$

❑ *Estimar as frequências esperadas:*

$$E_{ijk} = \frac{O_{i..} \times O_{.j.} \times O_{..k}}{n^2} \Leftrightarrow E_{111} = \frac{O_{1..} \times O_{.1.} \times O_{..1}}{n^2} = \frac{483 \times 428 \times 110}{385^2} = 32{,}61$$

$$\vdots$$

$$E_{225} = \frac{O_{2..} \times O_{.2.} \times O_{..5}}{n^2} = \frac{352 \times 407 \times 154}{385^2} = 31{,}64$$

Tabela da frequência esperada

Gén.	P.P.	Ideologia política				
		ML	LL	Mod	LC	MC
M	D	32,61	37,65	94,58	37,06	45,66
	R	31,01	35,81	89,94	35,24	43,42
H	D	23,77	27,44	68,93	27,01	33,28
	R	22,60	26,10	65,55	25,68	31,64

99

- *Estatística do teste:*

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\frac{\left(O_{ijk} - E_{ijk}\right)^2}{E_{ijk}} = \frac{(44 - 32{,}61)^2}{32{,}61} + \cdots + \frac{(51 - 31{,}64)^2}{31{,}64} \approx 81{,}50$$

- *Grau de liberdade:* $gl = rcl - r - c - l + 2$
  $= 2 \times 2 \times 5 - 2 - 2 - 5 + 2$
  $= 13$

- *Ponto crítico:* $\alpha = 0{,}05 \Longrightarrow \chi^2_{(1-\alpha);gl} = \chi^2_{(0{,}95);24} \approx 22{,}36$

- *Decisão:* Como $\boldsymbol{\chi^2_{obs} > \chi^2_{0{,}95;24}}$, rejeita-se H_0 em favor de H_1. Conclui-se, ao nível significância de 5%, que o género, o partido político e a ideologia política estão significativamente associados.

100

Exercício

Baseado no problema do exemplo, testar se:

1. Cada par de variáveis está associado com a terceira (independência parcial)
2. Duas variáveis estão associados para uma categoria específica da terceira (independência condicional)
3. As variáveis estão associados 2 a 2

Soluções:

1. (BC, A) $\Longrightarrow \chi^2_{obs} = 17{,}34$
   (AC, B) $\Longrightarrow \chi^2_{obs} = 67{,}37$
   (AB, C) $\Longrightarrow \chi^2_{obs} = 74{,}24$
2. (AB, AC) $\Longrightarrow \chi^2_{obs} = 62{,}42$
   (AB, BC) $\Longrightarrow \chi^2_{obs} = 12{,}13$
   (AC, BC) $\Longrightarrow \chi^2_{obs} = 6{,}77$
3. (AB, AC, BC) $\Longrightarrow \chi^2_{obs} = 3{,}23$

101

4.2. Modelos log-lineares

À semelhança do efetuado para tabelas bidimensionais, nas tabelas tridimensionais devem ser estabelecidos:

Considere-se as seguintes notações:

$$\bar{n}_{...} = \frac{1}{r \times c \times l} \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{l} \ln E_{ijk}$$

$$\bar{n}_{i..} = \frac{1}{j \times k} \sum_{j=1}^{c} \sum_{k=1}^{l} \ln E_{ijk} \qquad \bar{n}_{ij.} = \frac{1}{k} \sum_{k=1}^{l} \ln E_{ijk}$$

$$\bar{n}_{.j.} = \frac{1}{i \times k} \sum_{i=1}^{r} \sum_{k=1}^{l} \ln E_{ijk} \qquad \bar{n}_{i.k} = \frac{1}{j} \sum_{j=1}^{c} \ln E_{ijk}$$

$$\bar{n}_{..k} = \frac{1}{i \times j} \sum_{i=1}^{r} \sum_{j=1}^{c} \ln E_{ijk} \qquad \bar{n}_{.jk} = \frac{1}{i} \sum_{i=1}^{r} \ln E_{ijk}$$

$\mu = \bar{n}_{...}$; efeito médio global

$\lambda_i^A = \bar{n}_{i..} - \mu$; efeito da variável A

$\lambda_j^B = \bar{n}_{.j.} - \mu$; efeito da variável B

$\lambda_k^C = \bar{n}_{..k} - \mu$; efeito da variável B

$\lambda_{ij}^{AB} = \bar{n}_{ij.} - \bar{n}_{i..} - \bar{n}_{.j.} + \mu$; efeito da intersecção entre A e B

$\lambda_{ik}^{AC} = \bar{n}_{i.k} - \bar{n}_{i..} - \bar{n}_{..k} + \mu$; efeito da intersecção entre A e C

$\lambda_{jk}^{BC} = \bar{n}_{.jk} - \bar{n}_{.j.} - \bar{n}_{..k} + \mu$; efeito da intersecção entre B e C

$\lambda_{ijk}^{ABC} = \ln E_{ijk} - \bar{n}_{ij.} - \bar{n}_{i.k} - \bar{n}_{.jk} - \bar{n}_{i..} - \bar{n}_{.j.} - \bar{n}_{..k} - \mu$; efeito da intersecção entre A, B e C ou intersecção da 2ª ordem

Restrição adicional:

$$\sum_{i=1}^{r} \lambda_i^A = \sum_{j=1}^{c} \lambda_j^B = \sum_{k=1}^{l} \lambda_k^C = \sum_{i=1}^{r} \lambda_{ij}^{AB} = \cdots = \sum_{j=1}^{c} \lambda_{ijk}^{ABC} = \sum_{k=1}^{l} \lambda_{ijk}^{ABC} = 0$$

102

a) Modelo saturado (modelo completo)

Na tabela bidimensional:

$$\ln E_{ij} = \underbrace{\mu}_{\substack{efeito \\ constante}} + \underbrace{\lambda_i^A + \lambda_j^B}_{\substack{efeitos\ isolados \\ de\ A,B}} + \underbrace{\lambda_{ij}^{AB}}_{dependência}$$

Na tabela tridimensional:

$$\ln E_{ijk} = \underbrace{\mu}_{\substack{efeito \\ constante}} + \underbrace{\lambda_i^A + \lambda_j^B + \lambda_k^C}_{\substack{efeitos\ isolados \\ de\ A,B,C}} + \underbrace{\lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}}_{dependência}$$

- $\lambda_{ij}^{AB}, \lambda_{ik}^{AC}, \lambda_{jk}^{BC}$ são os efeitos de dependência 2 a 2
- λ_{ijk}^{ABC} é o efeito de dependência conjunto de A, B, C

103

b) Modelo independência

- Associação 2 a 2 (AB, AC, BC)

$$\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

- Independência condicional $\{(AC, BC), (AB, BC), (AB, AC)\}$:
  - $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} \longrightarrow$ independência condicional entre A e B dada C
  - $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} \longrightarrow$ independência condicional entre A e C dada B
  - $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} \longrightarrow$ independência condicional entre B e C dada A
- Independência parcial $\{(AB, C), (AC, B), (BC, A)\}$:
  - $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} \longrightarrow$ independência parcial entre (AB) e C
  - $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} \longrightarrow$ independência parcial entre (AC) e B
  - $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{jk}^{BC} \longrightarrow$ independência parcial entre (BC) e A
- Independência mútua (A, B, C): $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C$

104

Ajustamento de modelos log-lineares

1) Informação mínimas para estimar E_{ijk} nos modelos log-linear abrangentes

Modelo		**Símbolo**	**Informação***
Saturado/completo		(ABC)	$\{O_{ijk}\}$
Associação 2 a 2		(AB, AC, BC)	$\{O_{ij.}\}, \{O_{i.k}\}, \{O_{.jk}\}$
Independência condicional	entre B e C, dado A	(AB, AC)	$\{O_{ij.}\}, \{O_{i.k}\}$
	entre A e C, dado B	(AB, BC)	$\{O_{ij.}\}, \{O_{.jk}\}$
	entre A e B, dado C	(AC, BC)	$\{O_{i.k}\}, \{O_{.jk}\}$
Independência parcial	entre (B,C) e A	(BC, A)	$\{O_{.jk}\}, \{O_{i..}\}$
	entre (A,C) e B	(AC, B)	$\{O_{i.k}\}, \{O_{.j.}\}$
	entre (A, B) e C	(AB, C)	$\{O_{ij.}\}, \{O_{..k}\}$
Independência mútua		(A, B, C)	$\{O_{i..}\}, \{O_{.j.}\}, \{O_{..k}\}$

*Esta informação é a necessária para obter os valores de E_{ijk} para cada modelo considerado (constitui as chamadas "Estatísticas suficientes mínimas"). Por exemplo, o mínimo de informação que é "suficiente" para obter E_{ijk} no modelo de independência mútua é $O_{i..}$, $O_{.j.}$ e $O_{..k}$.

105

2) Estimar as frequências esperadas dos modelos

Modelo		Símbolo	$\widehat{E}_{ijk}$
Saturado/completo		(ABC)	O_{ijk}
Associação 2 a 2		(AB, AC, BC)	Método iterativo
Independência condicional	entre B e C, dado A	(AB, AC)	$\frac{O_{ij.} \times O_{i.k}}{O_{i..}}$
	entre A e C, dado B	(AB, BC)	$\frac{O_{ij.} \times O_{.jk}}{O_{.j.}}$
	entre A e B, dado C	(AC, BC)	$\frac{O_{i.k} \times O_{.jk}}{O_{..k}}$
Independência parcial	entre (B,C) e A	(BC, A)	$\frac{O_{.jk} \times O_{i..}}{n}$
	entre (A,C) e B	(AC, B)	$\frac{O_{i.k} \times O_{.j.}}{n}$
	entre (A, B) e C	(AB, C)	$\frac{O_{ij.} \times O_{..k}}{n}$
Independência mútua		(A, B, C)	$\frac{O_{i..} \times O_{.j.} \times O_{..k}}{n^2}$

106

3) Ajuste do modelo

Para a testar o ajustamento global dos modelos log-lineares usa-se a estatística qui-quadrado (χ^2) ou a estatística de razão de verosimilhança (G^2) dados por

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}} \quad \text{e} \quad G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\left(O_{ijk}\ln\frac{O_{ijk}}{E_{ijk}}\right).$$

Quando o tamanho de amostra é elevado χ^2 e G^2 têm aproximadamente uma distribuição Qui-quadrado com $\boldsymbol{\nu}$ graus de liberdade (onde $\boldsymbol{\nu}$ dependendo do modelo considerado).

4) Seleção de modelos

Considere-se, dois modelos $\boldsymbol{M_a}$ com $\boldsymbol{\nu_a}$ e $\boldsymbol{M_b}$ com $\boldsymbol{\nu_b}$ graus de liberdade. Sendo $\boldsymbol{M_a}$ um caso particular de $\boldsymbol{M_b}$, então $\boldsymbol{M_a}$ e $\boldsymbol{M_b}$ dizem-se modelos "encaixados". A comparação deste tipo de modelos pode ser feita mediante o cálculo da estatística:

$$G^2(M_a|M_b) = G^2(M_a) - G^2(M_b) \sim \chi^2_{(|\boldsymbol{\nu_b}-\boldsymbol{\nu_a}|)}$$

Decisão: Rejeita-se H_0 (em favor de H_1) quando $G^2(M_a|M_b) \geq \chi^2_{\alpha;(|\boldsymbol{\nu_b}-\boldsymbol{\nu_a}|)}$.

Se $G^2(M_a|M_b) < \chi^2_{\alpha;(|\boldsymbol{\nu_b}-\boldsymbol{\nu_a}|)}$ então não se rejeita H_0.

107

Modelos encaixados

Por exemplo,

Modelo saturado: $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$

Associação 2 a 2: $\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$

Permite testar a significância da interação conjunta de três variáveis

Hipóteses a testar:

$H_0: \lambda_{ijk}^{ABC} = 0$ vs. $H_1: \lambda_{ijk}^{ABC} \neq 0$

A rejeição de H_0 leva a concluir que o modelo saturado é aquele que descreve adequadamente os dados. Caso contrário, deve-se considerar o modelo "Associação 2 a 2" ou um outro modelo encaixado nele.

108

Se $H_0: \lambda_{ijk}^{ABC} = 0$ não for rejeitado, testa-se adicionalmente,

Independência condicional

- (AB, AC, BC) vs. (AB, AC) $H_0: \lambda_{.jk}^{BC} = 0$ vs. $H_1: \lambda_{.jk}^{BC} \neq 0$
- (AB, AC, BC) vs. (AB, BC) $H_0: \lambda_{i.k}^{AC} = 0$ vs. $H_1: \lambda_{i.k}^{AC} \neq 0$
- (AB, AC, BC) vs. (AC, BC) $H_0: \lambda_{ij.}^{AB} = 0$ vs. $H_1: \lambda_{ij.}^{AB} \neq 0$

(1) Se as 3 hipóteses H_0 forem rejeitadas então o modelo sem interação conjunta de três variáveis (i.e. o modelo "associação 2 a 2") é adequado.

(2) A não rejeição de uma destas hipóteses H_0 implica a independência mútua entre o respetivo par de variáveis. Assim, a não rejeição destas 3 H_0 simultaneamente implica a independência mútua entre as 3 variáveis. Então, o modelo de independência mútua é o adequado.

(3) Nas situações intermédias é ainda necessário testar a independência parcial:

- (AB, AC, BC) vs. (AB, C) $H_0: \lambda_{i.k}^{AC} = \lambda_{.jk}^{BC} = 0$ vs. $H_1: \sim H_0$
- (AB, AC, BC) vs. (AC, B) $H_0: \lambda_{ij.}^{AB} = \lambda_{.jk}^{BC} = 0$ vs. $H_1: \sim H_0$
- (AB, AC, BC) vs. (BC, A) $H_0: \lambda_{ij.}^{AB} = \lambda_{i.k}^{AC} = 0$ vs. $H_1: \sim H_0$

109

Exemplo

Retorne-se ao problema do exemplo anterior (slide 32), e identifique qual é o modelo log-linear mais adequado.

Género	Partido político	Ideologia política				
		Muito liberal	Levemente liberal	Moderada	Levemente conservadora	Muito conservadora
M	Democrata	44	47	118	23	32
	Republicano	18	28	86	39	48
H	Democrata	36	34	53	18	23
	Republicano	12	18	62	45	51

110

Resolução do exemplo

❑ *Os totais marginais*

- *De uma só variável:*

O género	
M	483
H	352

O part. pol.	
D	428
R	407

A ideologia política				
ML	LL	mod	LC	MC
110	127	319	125	154

- *De duas variáveis:*

Gén.	P.P.	
	D	R
M	264	219
H	164	188

Gén	Ideologia política				
	ML	LL	mod	LC	MC
M	62	75	204	62	80
H	48	52	115	63	74

P.P.	Ideologia política				
	ML	LL	mod	LC	MC
D	80	81	171	41	55
R	30	46	148	84	99

111

❑ Comparação dos modelos:

- Modelo (ABC) vs. Modelo (AB,AC,BC)

Hipóteses a testar:

H_0: associação 2 a 2 (M_a) vs. H_1: saturado (M_b)

Tabela de freq. Esp. do modelo (ABC)

Gén.	P.P.	Ideologia política				
		ML	LL	Mod	LC	MC
M	D	44	47	118	23	32
	R	18	28	86	39	48
H	D	36	34	53	18	23
	R	12	18	62	45	51

$G^2 = 53{,}29 \ ; \ gl = 0$

Tabela de Freq. Esp. do modelo (AB,AC,BC)

Gén.	P.P.	Ideologia política				
		ML	LL	Mod	LC	MC
M	D	46,58	49,80	114,40	22,23	30,99
	R	15,42	25,20	89,60	39,77	49,01
H	D	33,42	31,20	56,60	18,77	24,01
	R	14,58	20,80	58,40	44,23	49,99

$G^2 = 32{,}45 \ ; \ gl = 4$

$G^2(M_a|M_b) = G^2(M_a) - G^2(M_b) = 32{,}45 - 53{,}29 = -20{,}84$

Ponto crítico: $\alpha = 0{,}05; \Leftrightarrow \chi^2_{\alpha;(|\nu_b - \nu_a|)} = \chi^2_{0{,}05;4} = 9{,}49$

Como $G^2(M_a|M_b) > \chi^2_{0{,}05;4}$, concluirmos rejeitar H_0 em favor de H_1

112

- Modelo (AB,AC,BC) vs. Modelo independência condicional

E_{ijk} do modelo (AB,AC)

G.	P.P.	Ideologia política				
		ML	LL	Mod	LC	MC
M	D	33,89	40,99	111,50	33,89	43,73
	R	28,11	34,01	92,50	28,11	36,27
H	D	22,36	24,23	53,58	29,35	34,48
	R	25,64	27,77	61,42	33,65	39,52

$G^2(M_a) = 63{,}80 \ ; \ gl = 8$

E_{ijk} do modelo (AB,BC)

G.	P.P.	Ideologia política				
		ML	LL	Mod	LC	MC
M	D	49,35	49,96	105,48	25,29	33,93
	R	16,14	24,75	79,64	45,20	53,27
H	D	30,65	31,04	65,52	15,71	21,07
	R	13,86	21,25	68,36	38,80	45,73

$G^2(M_a) = 12{,}21 \ ; \ gl = 8$

E_{ijk} do modelo (AC,BC)

G.	P.P.	Ideologia política				
		ML	LL	Mod	LC	MC
M	D	45,09	47,83	109,35	20,34	28,57
	R	16,91	27,17	94,65	41,66	51,43
H	D	34,91	33,17	61,65	20,66	26,43
	R	13,09	18,83	53,35	42,34	47,57

$G^2(M_a) = 67{,}76 \ ; \ gl = 5$

- $G^2(M_a|M_b) = G^2(M_a) - G^2(M_b) = 63{,}80 - 32{,}45 = 31{,}35$

  Ponto crítico: $\alpha = 0{,}05; \Leftrightarrow \chi^2_{\alpha;(|\nu_b - \nu_a|)} = \chi^2_{0{,}05;4} = 9{,}49$
- $G^2(M_a|M_b) = G^2(M_a) - G^2(M_b) = 12{,}21 - 32{,}45 = -20{,}24$

  Ponto crítico: $\alpha = 0{,}05; \Leftrightarrow \chi^2_{\alpha;(|\nu_b - \nu_a|)} = \chi^2_{0{,}05;4} = 9{,}49$
- $G^2(M_a|M_b) = G^2(M_a) - G^2(M_b) = 74{,}54 - 32{,}45 = 42{,}09$

  Ponto crítico: $\alpha = 0{,}05; \Leftrightarrow \chi^2_{\alpha;(|\nu_b - \nu_a|)} = \chi^2_{0{,}05;1} = 3{,}84$

<u>Conclusão</u>:

Como há hipóteses H_0 rejeitadas e não rejeitadas então é ainda necessário testar a independência parcial.

113

- Modelo (AB,AC,BC) vs. Modelo independência parcial

E_{ijk} do modelo (BC,A)

G.	P.P.	Ideologia política				
		ML	LL	Mod	LC	MC
M	D	46,28	46,85	98,91	23,72	31,81
	R	17,35	26,61	85,61	48,59	57,27
H	D	33,72	34,15	72,09	17,28	23,19
	R	12,65	19,39	62,39	35,41	41,73

$G^2(M_a) = 17{,}52;\ gl = 9$

E_{ijk} do modelo (AC,B)

G.	P.P.	Ideologia política				
		ML	LL	Mod	LC	MC
M	D	31,78	38,44	104,57	31,78	41,01
	R	30,22	36,56	99,43	30,22	38,99
H	D	24,60	26,65	58,95	32,29	50,74
	R	23,40	25,35	56,05	30,71	36,07

$G^2(M_a) = 69{,}11\ ;\ gl = 9$

E_{ijk} do modelo (AB,C)

G.	P.P.	Ideologia política				
		ML	LL	Mod	LC	MC
M	D	34,78	40,15	100,86	39,52	48,69
	R	28,85	33,31	83,67	32,78	40,39
H	D	21,60	24,94	62,65	24,55	30,25
	R	24,77	28,59	71,82	28,14	34,67

$G^2(M_a) = 74{,}54\ ;\ gl = 12$

- $G^2(M_a|M_b) = G^2(M_a) - G^2(M_b) = 17{,}52 - 32{,}45 = -14{,}93$

  Ponto crítico: $\alpha = 0{,}05; \Leftrightarrow \chi^2_{\alpha;(|\nu_b - \nu_a|)} = \chi^2_{0{,}05;5} = 11{,}07$
- $G^2(M_a|M_b) = G^2(M_a) - G^2(M_b) = 69{,}11 - 32{,}45 = 36{,}66$

  Ponto crítico: $\alpha = 0{,}05; \Leftrightarrow \chi^2_{\alpha;(|\nu_b - \nu_a|)} = \chi^2_{0{,}05;5} = 11{,}07$
- $G^2(M_a|M_b) = G^2(M_a) - G^2(M_b) = 74{,}54 - 32{,}45 = 42{,}09$

  Ponto crítico: $\alpha = 0{,}05; \Leftrightarrow \chi^2_{\alpha;(|\nu_b - \nu_a|)} = \chi^2_{0{,}05;8} = 15{,}51$

Conclusão:
Como há hipóteses H_0 rejeitadas e não rejeitadas então é ainda necessário fazer a seleção para os modelos que H_0 não é rejeitada.

114

- O modelo adequado

O modelo adequado é escolhido dos modelos para os que H_0 não é rejeitada, em comparações dos modelos encaixados. Considera-se o que tem menor grau de liberdade.

Modelo	gl	G^2	X^2
(AB,AC,BC)	4	32,29	32,35
(AB,BC)	8	12,21	12,13
(BC,A)	9	17,52	17,34

Logo, o modelo adequado é o modelo associação 2 a 2, isto é (AB,AC,BC)

$$\ln E_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

NOTA:
O modelo adequado é aquele que, apresentado um bom ajustamento aos dados, seja interpretável e possua mínimo número de parâmetros (o mais parcimonioso).

115

- Estimar os parâmetros do modelo

μ	λ_i^A	λ_j^B	λ_k^C	$\lambda_{ij.}^{AB}$	$\lambda_{i.k}^{AC}$	$\lambda_{.jk}^{BC}$
3,84	$\lambda_1^A = 0$	$\lambda_1^B = 0$	$\lambda_1^C = 0$	$\lambda_{11.}^{AB} = 0$	$\lambda_{1.1}^{AC} = 0$	$\lambda_{.11}^{BC} = 0$
	$\lambda_2^A = -0{,}33$	$\lambda_2^B = -1{,}11$	$\lambda_2^C = 0{,}07$	$\lambda_{12.}^{AB} = 0$	$\lambda_{1.2}^{AC} = 0$	$\lambda_{.12}^{BC} = 0$
			$\lambda_3^C = 0{,}90$	$\lambda_{21.}^{AB} = 0$	$\lambda_{1.3}^{AC} = 0$	$\lambda_{.13}^{BC} = 0$
			$\lambda_4^C = -0{,}74$	$\lambda_{22.}^{AB} = 0{,}28$	$\lambda_{1.4}^{AC} = 0$	$\lambda_{.14}^{BC} = 0$
			$\lambda_5^C = -0{,}41$		$\lambda_{1.5}^{AC} = 0$	$\lambda_{.15}^{BC} = 0$
					$\lambda_{2.1}^{AC} = 0$	$\lambda_{.21}^{BC} = 0$
					$\lambda_{2.2}^{AC} = -0{,}14$	$\lambda_{.22}^{BC} = 0{,}42$
					$\lambda_{2.3}^{AC} = -0{,}37$	$\lambda_{.23}^{BC} = 0{,}86$
					$\lambda_{2.4}^{AC} = 0{,}16$	$\lambda_{.24}^{BC} = 1{,}69$
					$\lambda_{2.5}^{AC} = 0{,}08$	$\lambda_{.25}^{BC} = 1{,}56$

116

Bibliografia

1. Mello, F.M., 2014. *Dicionário de Estatística. 673 entradas Índice remissivo em Português e inglês. Edições Sílabo, Lisboa*
2. Howell, D. C. (2000). Chi-Square Test - Analysis of Contingency Tables, 1–4.
3. Reis, E., Melo, P., Andrade, R., & Calapez, T. (2016). *Estatística aplicada 2*. (Edições Sílabo, Ed.) (5a Edição). Lisboa.
4. Análise de Resíduos - Tabela Cruzada | Portal Action. (n.d.). Retrieved August 7, 2017, from http://www.portalaction.com.br/tabela-de-contingencia/analise-de-residuos
5. Leal, M. M. (1997). *Modelos Log-lineares em Tabelas de Contingência*. Lisboa.
6. Murteira, B., & Antunes, M. (2012). *Probabilidade e Estatística*. (Escolar Editora, Ed.) (Volume II). Lisboa.

Folha prática 2

Universidade Nacional de Timor Lorosa'e (UNTL)
Faculdade de Ciências Exatas (FCE)
Estatística e Análise de Dados ano letivo 2019, 1º semestre

Análise de tabelas de contingência

1. A tabela seguinte apresenta os resultados observados, numa amostra aleatória constituída por 564 participantes, expressos em frequências absolutas de acordo com a categoria socioprofissional e o meio de comunicação social preferido.

		Categoria socioprofissional				
		Agricultor	Quadro superior	Quadro média	Operário	**Totais**
Meio de comunicação social	Televisão	26	19	44	181	**270**
	Jornais	18	49	87	107	**261**
	Rádio	9	4	4	16	**33**
	Totais	**53**	**72**	**135**	**304**	**564**

Nestas condições, responda às seguintes questões:

a) Qual é o meio de comunicação social mais importante para a população?

b) Qual é o meio de comunicação social mais importante para cada nível socioprofissional?

c) O que poderá concluir deste estudo? Será que o meio de comunicação social preferido é o mesmo para todas categorias socioprofissionais? Justifique.

d) Será que escolha do meio de comunicação social depende de categoria socioprofissional? Caso existe a dependência, responda as seguintes questões:

   i) Quantificar o seu grau de associação.

   ii) Identificar quais são os meios de comunicação escolhidos pela categoria socioprofissional, que mais contribuem para a dependência.

e) Determine o modelo log-linear adequado que se ajusta aos dados, quantifique os efeitos das variáveis e interprete os resultados obtidos.

2. Os alunos de uma dada disciplina, com 3 turmas, são classificados de acordo com a sua nota de teste em 3 níveis:

	< 10	**10 – 14**	**> 14**
Turma A	10	15	5
Turma B	20	10	5
Turma C	10	15	10

Determine:

a) A probabilidade de um aluno a realizar o teste ser da turma A.
b) A probabilidade de um aluno ter nota inferior de 10.
c) Qual é a percentagem dos alunos a realizar o teste por cada turma.
d) A probabilidade de um aluno ser da turma B e ter nota no intervalo 10 a 14.
e) A probabilidade de um aluno da turma C ter nota superior de 14.
f) Será que o desempenho escolar (notas obtidas) é o mesmo para as três turmas?
g) Construa o modelo log-linear adequado; quantifique os efeitos das variáveis do modelo e interprete os resultados obtidos.

3. O número de transações efetuadas em três moedas estrangeiras por duas empresas nacionais, num total de 1000 transações, foi registado durante um longo período de tempo de acordo com a tabela que se segue:

	Empresa 1	Empresa 2
DEM	110	90
GBP	120	180
FRF	350	150

DEM: Marco Alemão; GBP: Libra Inglesa; FRF: Franco Francês

Tendo sido selecionada aleatoriamente uma transação de entre as registradas, diga qual a probabilidade de:

a) Ter sido realizada em DEM?
b) Ter sido realizada pela empresa 1?
c) Ter sido realizada pela empresa 1 ou em FRF?
d) Sabendo que foi realizada em GBP, ter sido realizada pela empresa 1?
e) Sabendo que não foi realizada pela empresa 1, não ter sido realizada em GBP?
f) Será que a transação em três moedas estrangeiras depende da empresa?
g) Será que empresa 1 e empresa 2 têm a mesma distribuição de transações por moeda estrangeira?

4. Num determinado jogo de campeonato do mundo de futebol estiveram presentes 50000 espectadores, dos quais 25000 eram dinamarqueses, 20000 eram brasileiros e 23000 eram mulheres. Dos dinamarqueses 12000 eram homens, enquanto dos Brasileiros 8000 eram mulheres. Com base nestes dados diga qual a probabilidade de:

a) Selecionado aleatoriamente um espectador, que o mesmo não seja de nacionalidade brasileira nem dinamarquesa?
b) Selecionando aleatoriamente um espectador, que o mesmo seja um homem.
c) Tendo-se selecionado aleatoriamente um espectador e verificado que era uma mulher, seja de nacionalidade brasileira?

d) Tendo-se selecionada aleatoriamente um espectador e verificado que era um homem, seja de nacionalidade brasileira?

e) Tendo-se selecionado aleatoriamente um espectador e verificado que não era uma mulher, seja de nacionalidade dinamarquesa?

5. Existem três modos de efetuar os pagamentos num supermercado durante o período do dia, são: por cheque, dinheiro e cartão de crédito. A seguinte tabela de contingência apresenta os resultados obtidos numa amostra de 4000 clientes:

	Período do dia		
Modo de pagamento	Manhã	Tarde	Noite
Cheque	750	1500	750
Dinheiro	125	300	75
Cartão de crédito	125	200	175

a) Testar ao nível de significância de 5% a hipótese de que o modo de pagamento dos clientes nesse estabelecimento é independente do período do dia em que fazem as compras.

b) Avaliar também se existe homogeneidade entre os períodos do dia com categoria de modo de pagamentos que distribuem pelas dinheiro e cartão credito.

6. Pretende-se avaliar se a reprodução dos melros está associada ao tipo de habitat. Foram marcados 30 ninhos no habitat agrícola e 35 no habitat florestal. Dos 30 foram bem-sucedidos 20 e dos 35 foram bem-sucedidos 10.

a) Indique as hipóteses em teste.

b) Efetue o teste adequado.

7. Para se verificar se o apoio à instalação de uma nova avenida dependia do local onde as pessoas moram, foram entrevistados 1000 moradores, igualmente divididos em quatro diferentes bairros. Podemos afirmar, com 5% de significância, que a opinião dos cidadãos depende do bairro onde eles moram?

	Favoráveis	Contrários
A	130	120
B	125	125
C	165	85
D	160	90

8. A seguinte tabela apresenta os resultados do lançamento de um dado lançado 1000 vezes. Será que os resultados obtidos sustentam a hipótese de que o dado é honesto?

x_i	O_i
1	174
2	174
3	154
4	179
5	154
6	165
Total	1000

9. A procura diária de um certo produto foi, em 40 dias escolhidos ao acaso, a seguinte:

Número de unidades	Número de dias
0	6
1	14
2	10
3	7
4	2
5	1

Será que tais observações foram extraídas de uma população com distribuição Poisson, isto é, será de admitir que tal procura segue uma distribuição de Poisson?

10. Em 100 lançamentos de uma moeda, observaram-se 65 coroas e 35 caras. Testar a hipótese de a moeda ser honesta, adotando-se $\alpha = 5\%$.

11. Deseja-se verificar se o número de acidentes numa estrada muda conforme o dia da semana. O número de acidentes observado para cada dia de uma semana escolhida aleatoriamente foram:

Dia da semana	Nº.de acidentes
Seg.	20
Ter.	10
Qua.	10
Qui.	15
Sex.	30
Sáb.	20
Dom.	35

12. Pretende-se construir um modelo de simulação das operações de um determinado terminal de um porto situado na Europa. Uma das variáveis a considerar no modelo é a diferença entre a data de chegada dos navios provenientes dos EU e a respetiva data planeada. Dado que tal diferença é influenciada por muitos fatores, pode considerar-se como uma variável aleatória, e pretende-se testar se esta diferença segue uma distribuição normal de média 0,1 e desvio padrão 7,2. Uma amostra de 30 navios revelou os resultados são seguintes:

-6,6	-2	5	2,4	-1,8	-0,3	15	-7,6	-0,6	2,6
-7,4	12,4	-6	-5,8	15,2	-2,4	-8,9	-5,6	-3,7	2,2
8,2	-9	13,2	7,6	-2,8	-1,8	1,8	4,4	2,2	4

13. Os dados a seguir registam o tratamento mortalidade para 22 doentes de AIDS

Tratamento	**Mortalidade**		**total**
	Sim	**Não**	
A	7	5	12
B	1	9	10
total	8	14	22

Testar se existe associação entre tipo de tratamento e mortalidade por AIDS.

14. Duas medicações, A e B, foram avaliadas em 100 doentes com cefaleias frequentes. Foi-lhes pedido para tomarem o medicamento A durante um mês e, no mês seguinte, tomarem o medicamento B. Pediu-se aos doentes que registassem se durante cada mês tiveram ou não dores de cabeça.

	A – s/cefaleias	A – c/cefaleias	total
B – s/cefaleias	45	4	49
B – c/cefaleias	17	34	51
total	62	38	100

a) Terá existido uma mudança significativa a dor de cabeça após de consumo dos medicamentos?
b) Qual a percentagem de doentes com cefaleias com o medicamento B? E para os doentes com cefaleias usando medicamento A?

15. Num estudo sobre relação entre o rendimento, a posição socio-Professional dos pais e nível de educação de um grupo de jovens dinamarqueses (com idade compreendidas entre os 14 e os 20 anos) obteve-se a seguinte tabela de contingência:

Nível soc.	Educação	Rendimentos			
		1	2	3	4
1	1	9	21	62	104
	2	2	8	15	59
	3	1	20	45	48
	4	2	8	13	16
2	1	46	228	126	39
	2	8	24	19	14
	3	49	187	132	43
	4	47	162	45	12
3	1	30	141	86	10
	2	2	13	7	6
	3	45	171	104	13
	4	38	102	29	5
4	1	33	170	67	4
	2	1	5	0	2
	3	40	176	81	3
	4	54	181	55	2

**Os níveis socio-profissional:*

1 – Dono de uma grande ou média empresa ou empregado com posição de chefia;

2 – Dono de uma pequena empresa ou empregado especializado;

3 – Dono de uma empresa muito pequena ou seu empregado especializado;

4 – Trabalhador não especializado.

**Os Níveis de rendimento:*

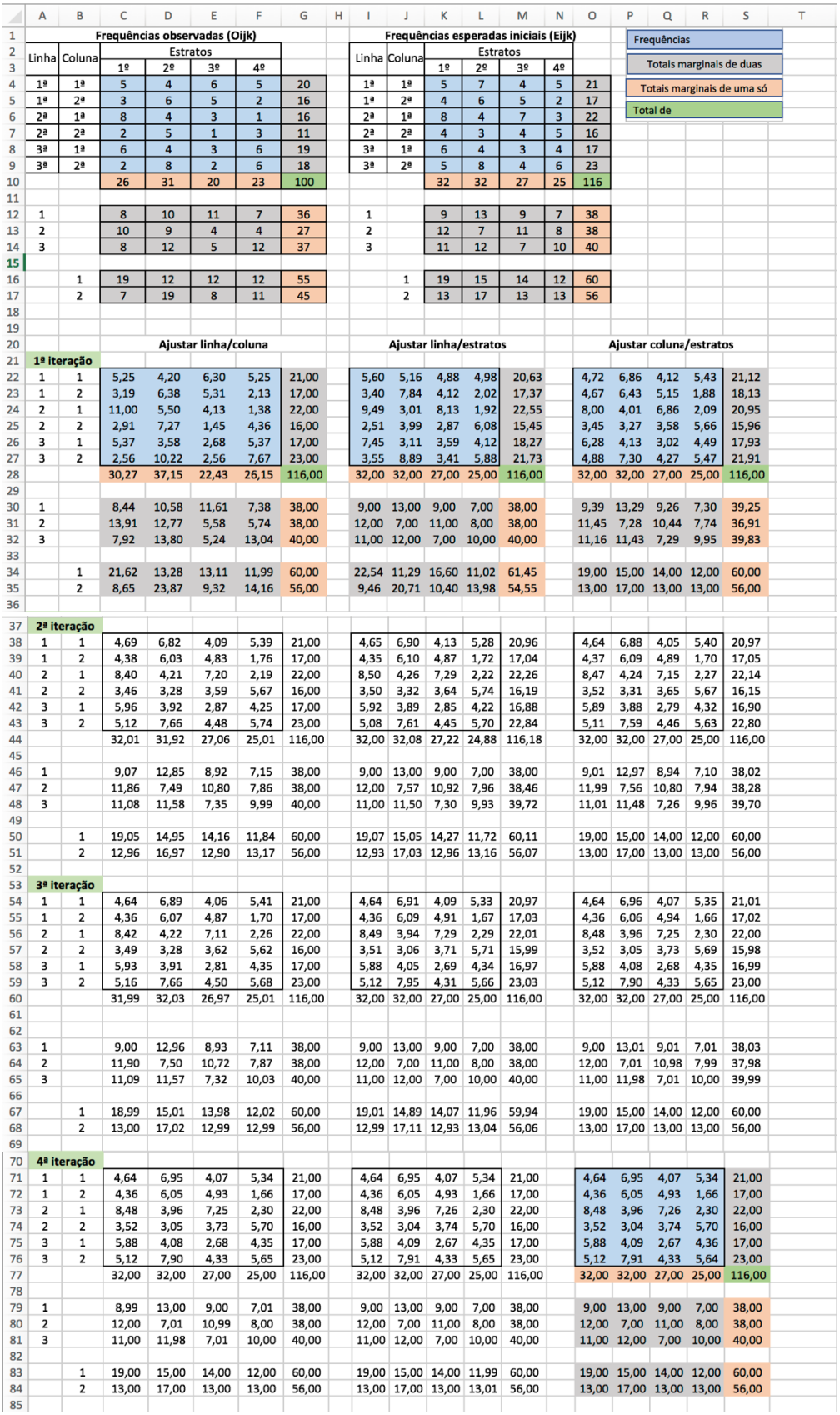
1 : 0 – 999; 2: 1000 – 1999; 3: 2000 – 2999; 4: 3000 ou mais.

**Os níveis da Educação*

1 – Ainda estudar; 2 – Licenciado; 3 – Curso de formação Professional; 4 – Ensino escolar obrigatório.

a) Testar se existe associação entre o rendimento, a posição socio-Professional a educação de um jovem.
b) Quantificar o grau de dependência e, caso exista dependência, identificar quais são as variáveis que mais contribuem para a dependência.
c) Determine o modelo log-linear adequado que se ajusta aos dados, quantifique os efeitos das variáveis e interprete os resultados obtidos.

Folha de cálculo de excel para método iterativo



			Frequências observadas (Oijk)								Frequências esperadas iniciais (Eijk)															
	Linha	Coluna	Estratos						Linha	Coluna	Estratos															
			1º	2º	3º	4º					1º	2º	3º	4º												
	1ª	1ª	5	4	6	5	20		1ª	1ª	5	7	4	5	21											
	1ª	2ª	3	6	5	2	16		1ª	2ª	4	6	5	2	17											
	2ª	1ª	8	4	3	1	16		2ª	1ª	8	4	7	3	22											
	2ª	2ª	2	5	1	3	11		2ª	2ª	4	3	4	5	16											
	3ª	1ª	6	4	3	6	19		3ª	1ª	6	4	3	4	17											
	3ª	2ª	2	8	2	6	18		3ª	2ª	5	8	4	6	23											
			26	31	20	23	100				32	32	27	25	116											
	1		8	10	11	7	36		1		9	13	9	7	38											
	2		10	9	4	4	27		2		12	7	11	8	38											
	3		8	12	5	12	37		3		11	12	7	10	40											
		1	19	12	12	12	55			1	19	15	14	12	60											
		2	7	19	8	11	45			2	13	17	13	13	56											
			Ajustar linha/coluna								Ajustar linha/estratos								Ajustar coluna/estratos							
	1ª iteração																									
	1	1	5,25	4,20	6,30	5,25	21,00				5,60	5,16	4,88	4,98	20,63				4,72	6,86	4,12	5,43	21,12			
	1	2	3,19	6,38	5,31	2,13	17,00				3,40	7,84	4,12	2,02	17,37				4,67	6,43	5,15	1,88	18,13			
	2	1	11,00	5,50	4,13	1,38	22,00				9,49	3,01	8,13	1,92	22,55				8,00	4,01	6,86	2,09	20,95			
	2	2	2,91	7,27	1,45	4,36	16,00				2,51	3,99	2,87	6,08	15,45				3,45	3,27	3,58	5,66	15,96			
	3	1	5,37	3,58	2,68	5,37	17,00				7,45	3,11	3,59	4,12	18,27				6,28	4,13	3,02	4,49	17,93			
	3	2	2,56	10,22	2,56	7,67	23,00				3,55	8,89	3,41	5,88	21,73				4,88	7,30	4,27	5,47	21,91			
			30,27	37,15	22,43	26,15	116,00				32,00	32,00	27,00	25,00	116,00				32,00	32,00	27,00	25,00	116,00			
	1		8,44	10,58	11,61	7,38	38,00				9,00	13,00	9,00	7,00	38,00				9,39	13,29	9,26	7,30	39,25			
	2		13,91	12,77	5,58	5,74	38,00				12,00	7,00	11,00	8,00	38,00				11,45	7,28	10,44	7,74	36,91			
	3		7,92	13,80	5,24	13,04	40,00				11,00	12,00	7,00	10,00	40,00				11,16	11,43	7,29	9,95	39,83			
		1	21,62	13,28	13,11	11,99	60,00				22,54	11,29	16,60	11,02	61,45				19,00	15,00	14,00	12,00	60,00			
		2	8,65	23,87	9,32	14,16	56,00				9,46	20,71	10,40	13,98	54,55				13,00	17,00	13,00	13,00	56,00			
	2ª iteração																									
	1	1	4,69	6,82	4,09	5,39	21,00				4,65	6,90	4,13	5,28	20,96				4,64	6,88	4,05	5,40	20,97			
	1	2	4,38	6,03	4,83	1,76	17,00				4,35	6,10	4,87	1,72	17,04				4,37	6,09	4,89	1,70	17,05			
	2	1	8,40	4,21	7,20	2,19	22,00				8,50	4,26	7,29	2,22	22,26				8,47	4,24	7,15	2,27	22,14			
	2	2	3,46	3,28	3,59	5,67	16,00				3,50	3,32	3,64	5,74	16,19				3,52	3,31	3,65	5,67	16,15			
	3	1	5,96	3,92	2,87	4,25	17,00				5,92	3,89	2,85	4,22	16,88				5,89	3,88	2,79	4,32	16,90			
	3	2	5,12	7,66	4,48	5,74	23,00				5,08	7,61	4,45	5,70	22,84				5,11	7,59	4,46	5,63	22,80			
			32,01	31,92	27,06	25,01	116,00				32,00	32,08	27,22	24,88	116,18				32,00	32,00	27,00	25,00	116,00			
	1		9,07	12,85	8,92	7,15	38,00				9,00	13,00	9,00	7,00	38,00				9,01	12,97	8,94	7,10	38,02			
	2		11,86	7,49	10,80	7,86	38,00				12,00	7,57	10,92	7,96	38,46				11,99	7,56	10,80	7,94	38,28			
	3		11,08	11,58	7,35	9,99	40,00				11,00	11,50	7,30	9,93	39,72				11,01	11,48	7,26	9,96	39,70			
		1	19,05	14,95	14,16	11,84	60,00				19,07	15,05	14,27	11,72	60,11				19,00	15,00	14,00	12,00	60,00			
		2	12,96	16,97	12,90	13,17	56,00				12,93	17,03	12,96	13,16	56,07				13,00	17,00	13,00	13,00	56,00			
	3ª iteração																									
	1	1	4,64	6,89	4,06	5,41	21,00				4,64	6,91	4,09	5,33	20,97				4,64	6,96	4,07	5,35	21,01			
	1	2	4,36	6,07	4,87	1,70	17,00				4,36	6,09	4,91	1,67	17,03				4,36	6,06	4,94	1,66	17,02			
	2	1	8,42	4,22	7,11	2,26	22,00				8,49	3,94	7,29	2,29	22,01				8,48	3,96	7,25	2,30	22,00			
	2	2	3,49	3,28	3,62	5,62	16,00				3,51	3,06	3,71	5,71	15,99				3,52	3,05	3,73	5,69	15,98			
	3	1	5,93	3,91	2,81	4,35	17,00				5,88	4,05	2,69	4,34	16,97				5,88	4,08	2,68	4,35	16,99			
	3	2	5,16	7,66	4,50	5,68	23,00				5,12	7,95	4,31	5,66	23,03				5,12	7,90	4,33	5,65	23,00			
			31,99	32,03	26,97	25,01	116,00				32,00	32,00	27,00	25,00	116,00				32,00	32,00	27,00	25,00	116,00			
	1		9,00	12,96	8,93	7,11	38,00				9,00	13,00	9,00	7,00	38,00				9,00	13,01	9,01	7,01	38,03			
	2		11,90	7,50	10,72	7,87	38,00				12,00	7,00	11,00	8,00	38,00				12,00	7,01	10,98	7,99	37,98			
	3		11,09	11,57	7,32	10,03	40,00				11,00	12,00	7,00	10,00	40,00				11,00	11,98	7,01	10,00	39,99			
		1	18,99	15,01	13,98	12,02	60,00				19,01	14,89	14,07	11,96	59,94				19,00	15,00	14,00	12,00	60,00			
		2	13,00	17,02	12,99	12,99	56,00				12,99	17,11	12,93	13,04	56,06				13,00	17,00	13,00	13,00	56,00			
	4ª iteração																									
	1	1	4,64	6,95	4,07	5,34	21,00				4,64	6,95	4,07	5,34	21,00				4,64	6,95	4,07	5,34	21,00			
	1	2	4,36	6,05	4,93	1,66	17,00				4,36	6,05	4,93	1,66	17,00				4,36	6,05	4,93	1,66	17,00			
	2	1	8,48	3,96	7,25	2,30	22,00				8,48	3,96	7,26	2,30	22,00				8,48	3,96	7,26	2,30	22,00			
	2	2	3,52	3,05	3,73	5,70	16,00				3,52	3,04	3,74	5,70	16,00				3,52	3,04	3,74	5,70	16,00			
	3	1	5,88	4,08	2,68	4,35	17,00				5,88	4,09	2,67	4,35	17,00				5,88	4,09	2,67	4,36	17,00			
	3	2	5,12	7,90	4,33	5,65	23,00				5,12	7,91	4,33	5,65	23,00				5,12	7,91	4,33	5,64	23,00			
			32,00	32,00	27,00	25,00	116,00				32,00	32,00	27,00	25,00	116,00				32,00	32,00	27,00	25,00	116,00			
	1		8,99	13,00	9,00	7,01	38,00				9,00	13,00	9,00	7,00	38,00				9,00	13,00	9,00	7,00	38,00			
	2		12,00	7,01	10,99	8,00	38,00				12,00	7,00	11,00	8,00	38,00				12,00	7,00	11,00	8,00	38,00			
	3		11,00	11,98	7,01	10,00	40,00				11,00	12,00	7,00	10,00	40,00				11,00	12,00	7,00	10,00	40,00			
		1	19,00	15,00	14,00	12,00	60,00				19,00	15,00	14,00	11,99	60,00				19,00	15,00	14,00	12,00	60,00			
		2	13,00	17,00	13,00	13,00	56,00				13,00	17,00	13,00	13,01	56,00				13,00	17,00	13,00	13,00	56,00			

Frequências

Totais marginais de duas

Totais marginais de uma só

Total de

Atividades para aprendizagem com R

TESTE DE INDEPENDENCIA E HOMOGENEIDADE DE QUI-QUADRADO

```
#obs<-matrix(c(48,12,33,57,35,46,42,27),nr=2,nc=4,byrow=TRUE)
obs<-
 matrix(c(48,12,33,57,35,46,42,27),nr=2,nc=4,byrow=TRUE,dimnames=list(sexo=c("M",
 "H"), Cor_favorita=c("Preta","Branca","Vermelha","Azul")))
obs                          # frequências observadas
chisq.test(obs)              # teste Qui-quadrado
chisq.test(obs) $expected    # frequências esperadas
qchisq(0.95,3)               # Valor crítico, alfa = 0.05 (tabelado)
```

TESTE DE AJUSTAMENTO DE QUI-QUADRADO

```
Oij<-c(48,12,33,57)          # frequência observada
n<-sum(Oij);n                # dimensão da amostra
p<-rep(1/length(Oij),4);p    # probabilidade da distribuição uniforme (discreta)
Eij<-n*p;Eij                 # frequências esperadas
rbind(Oij,p,Eij)             # tabela de frequências observadas e esperadas
chisq.test(Oij,p=p)          # teste Qui-quadrado
chisq.test(Oij) $expected    # verificar as frequências esperadas

#x<-c(0,1,2,3,4,5)                      # observações
#Oij<-c(6,14,10,7,2,1)                  # frequências observadas
#n<-sum(Oij);n                          # dimensão da amostra
#media<-sum(x*Oij)/sum(Oij);media       # lambda (estimar parâmetro lambda)
#p<-dpois(x,media);p                    # f.d.p. da distribuição de Poisson
#Eij<-n*p;Eij                           # frequências esperadas
#chisq.test(Oij,p=p)                    # teste Qui-quadrado
#chisq.test(Oij) $expected              # verificar as frequências esperadas
```

TESTE EXATO DE FISHER

```
a<-matrix(c(1,3,3,2),2,2,byrow=TRUE);a
fisher.test(a,alternative="less")# alternative="less","greater" ou "two sided".
```

TESTE MCNEMAR

```
a<-matrix(c(14,7,26,33),2,2,byrow=TRUE);a
mcnemar.test(a,correct=FALSE)

b<-dbinom(3,16,.5)               # n=16, x=3, prob.= 0,5
round(b,2)
```

```
### MODELOS LOG-LINEARES ####

dados<- data.frame(expand.grid( # cria uma folha de dados de três colunas com
 todas as combinações dos níveis das três variáveis na sequencias C-B-A
 C=factor(c("ML","LL","Mod","LC","MC"),
 levels=c("ML","LL","Mod","LC","MC")),
 B=factor(c("Democrata","Republicano")),
 A=factor(c("Mulher","Homem"))),
contagem=c(44,47,118,23,32,18,28,86,39,48,36,34,53,18,23,12,18,62,45,51));dados

n<-sum(dados$contagem);n
#cat("\n tamanho da amostra:",n)

## Modelo (ABC)
m1<-glm(contagem~(B+C+A)^3, data=dados, family=poisson)
summary(m1)

## Modelo (AB,AC,BC)
m2<-glm(contagem~(B+C+A)^2, data=dados, family=poisson)
summary(m2)

## Modelo (AB,AC)
m31<-glm(contagem~B*A+C*A, data=dados, family=poisson)
summary(m31)

## Modelo (AB,BC)
m32<-glm(contagem~A*B+B*C, data=dados, family=poisson)
summary(m32)

## Modelo (AC,BC)
m33<-glm(contagem~A*C+B*C, data=dados, family=poisson)
summary(m33)

## Modelo (BC,A)
m41<-glm(contagem~B*C+A, data=dados, family=poisson)
summary(m41)

## Modelo (AC,B)
m42<-glm(contagem~A*C+B, data=dados, family=poisson)
summary(m42)

## Modelo (AB,C)
m43<-glm(contagem~A*B+C, data=dados, family=poisson)
summary(m43)

## Modelo (A,B,C)
m5<-glm(contagem~B+A+C,data=dados,family=poisson)
summary(m5)

## Graus de liberdade
gl<-
 c(m1$df.residual,m2$df.residual,m31$df.residual,m32$df.residual,m33$df.residual,
 m41$df.residual,m42$df.residual,m43$df.residual,m5$df.residual)
```

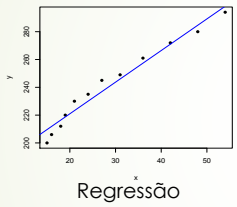
```
## G2 e p-valor
G2<-
 c(m1$deviance,m2$deviance,m31$deviance,m32$deviance,m33$deviance,m41$deviance,m4
 2$deviance,m43$deviance,m5$deviance)
PG2<-pchisq(G2,gl,lower.tail=FALSE)

## X2 e p-valor
X2<-
 c(sum(resid(m1,type="pearson")^2),sum(resid(m2,type="pearson")^2),sum(resid(m31,
 type="pearson")^2),sum(resid(m32,type="pearson")^2),sum(resid(m33,type="pearson"
 )^2),sum(resid(m41,type="pearson")^2),sum(resid(m42,type="pearson")^2),sum(resid
 (m43,type="pearson")^2),sum(resid(m5,type="pearson")^2))
PX2<-pchisq(X2,gl,lower.tail=FALSE)

## Bondade do ajuste
modelos<-
 c("(ABC)","(AB,AC,BC)","(AB,AC)","(AB,BC)","(AC,BC)","(BC,A)","(AC,B)","(AB,C)",
 "(A,B,C)")
print(data.frame(modelos,gl,G2,PG2,X2,PX2),digits=4)

## Frequências esperadas
freqest<-
 cbind(dados,fitted(m1),fitted(m2),fitted(m31),fitted(m32),fitted(m33),fitted(m41
 ),fitted(m42),fitted(m43),fitted(m5))
colnames(freqest)<-
 c("C","B","A","contagem","(ABC)","(AB,AC,BC)","(AB,AC)","(AB,BC)","(AC,BC)","(BC
 ,A)","(AC,B)","(AB,C)","(A,B,C)")
print(freqest,digits=4)
```

Capítulo 3 Regressão linear e análise de variância (ANOVA)

1

III. Regressão linear e análise de variância (ANOVA)

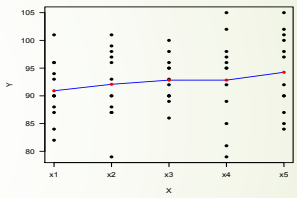
1. Introdução



Regressão

(x e y são quantitativos)

- $x_{ij}; i = 1,..,n; j = 1,...,k$ (fixos)
- n observações
- k regressores



ANOVA

(x qualitativo e y quantitativo)

- $X \rightarrow$ fator
- $x_j; j = 1,..,k$ (tratamentos/grupos)
- $x_j \rightarrow$ fixos ou aleatório
- $n \times k$ observações

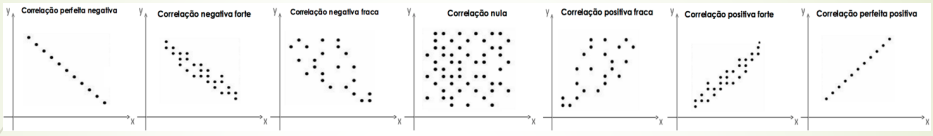
2

2. Correlação e regressão

Para analisar o relacionamento entre duas variáveis (uma variável independente e uma variável dependente) é possível construir um modelo matemático que melhor representa este relacionamento, e que é obtido por análise de regressão. Alternativamente, se o objetivo é simplesmente medir o grau ou força do relacionamento entre as variáveis pode ser realizada uma análise de correlação.

- uma variável dependente e uma variável independente ⟶ correlação simples,
- Uma variável dependente e mais do que uma variável independente ⟶ correlação múltipla.

3 A relação entre as variáveis pode ser observada através de um **diagrama de dispersão**



O grau de relacionamento linear pode ser quantificado através do ***coeficiente de correlação Pearson,*** definido por:

$$r = \frac{\overbrace{cov(x,y)}^{\text{variância conjunta entre } x \text{ e } y}}{\underbrace{\sqrt{var(x).var(y)}}_{\substack{\text{factor de normalização} \\ \text{que garante a propriedade} \\ -1 \le r \le 1}}} \Leftrightarrow r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Onde:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \Leftrightarrow S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \Leftrightarrow S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \Leftrightarrow S_{xy} = \sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y}$$

4 **Exemplo 1**

Considerar os dados sobre as horas de trabalho Y necessárias à produção de lotes de tamanho X, obtidos em 10 linhas de produção.

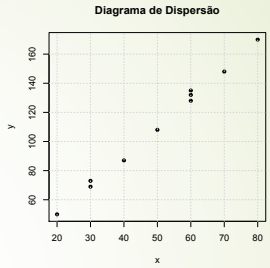
Produção	Tamanho de lote	Horas de trabalho
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

Averiguar se existe uma relação linear entre horas de trabalho (Y) e a produção de lotes de tamanho (X) através do diagrama de dispersão e quantificar o grau de relacionamento linear entre as variáveis.

5

Resolução do exemplo 1

Prod. (i)	Tam. de lote (x_i)	Horas de trab. (y_i)	x_i^2	y_i^2	$x_i y_i$
1	30	73	900	5329	2190
2	20	50	400	2500	1000
3	60	128	3600	16384	7680
4	80	170	6400	28900	13600
5	40	87	1600	7569	3480
6	50	108	2500	11664	5400
7	60	135	3600	18225	8100
8	30	69	900	4761	2070
9	70	148	4900	21904	10360
10	60	132	3600	17424	7920
Total	**500**	**1100**	**28400**	**134660**	**61800**



Na diagrama observamos que as horas de trabalho Y e a produção de lotes de tamanho X existe uma relação linear positiva forte.

- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{500}{10} = 50$ e $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1100}{10} = 110$
- $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 28400 - 10 \times 50^2 = 3400$
- $S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 134660 - 10 \times 110^2 = 13660$
- $S_{xy} = \sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y} = 61800 - 10 \times 50 \times 110 = 6800$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{6800}{\sqrt{34000 \times 13660}} \approx 0{,}998$$

6

3. Regressão linear simples

3.1. Modelo de Regressão Linear Simples (M.R.L.S.)

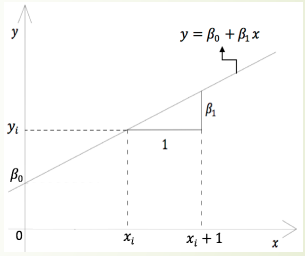
A relação do tipo linear entre duas variáveis pode ser descrita matematicamente através da equação $y = \beta_0 + \beta_1 x$, sendo:

y - variável explicada ou dependente

x - variável explicativa ou independente

β_0, β_1- parâmetros da regressão

- β_0 - coeficiente constante ou intercepto (ponto de intersecção da reta com eixo yy, isto é quando $x = 0$).
- β_1 - declive (ou coeficiente da regressão).



$y = \beta_0 + \beta_1 x$ ⟶ "reta de regressão" ou "reta ajustada"

Interpretação dos parâmetros do modelo

- β_0 - valor esperado para variável dependente y quando $x = 0$.
- β_1 - variação esperada na variável y, quando a variável x aumenta uma unidade.

7

3.2. Estimação e inferência sobre os parâmetros

a) Estimação dos parâmetros

Uma equação de regressão permite estimar valores de y, com base em valores conhecidos através de

$$\hat{y}_i = \beta_0 + \beta_1 x_i.$$

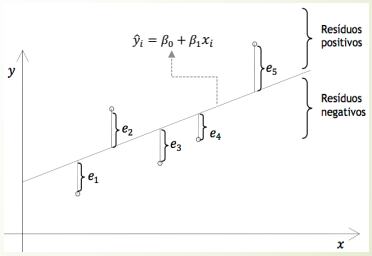
Assim, a relação entre y_i e x_i deve ser introduzir um termo $\boldsymbol{\varepsilon_i}$, que representa um *erro (resíduo) aleatório* que descreve o afastamento na vertical dos pontos em relação à reta de equação, isto é

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Leftrightarrow y_i = \hat{y}_i + \varepsilon_i$$

E o erro escreve-se

$$\varepsilon_i = y_i - \hat{y}_i \Leftrightarrow \varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

A menor distância entre os pontos (y) e a reta da regressão ($\hat{y}$), indica um melhor ajustamento da reta aos dados.



8

Portanto, as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ devem minimizar a soma dos quadrados dos erros.

$$\min_{\hat{\beta}_0, \hat{\beta}_1} S \text{ onde } S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Para tornar S mínima, é necessário derivar a função em ordem β_0 e β_1, igualar a zero e verificar o sinal da segunda derivada. Assim,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Onde, S_{xy} é a variância conjunta entre x e y, S_{xx} é a variância de x e $\bar{y}$ e $\bar{x}$ representam as médias de y e de x, respetivamente.

Assim, é possível estimar valores de y do seguinte modo:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

9 **Estimação dos parâmetros do modelo (Notação matricial)**

O modelo de regressão linear simples $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n$, pode ser escrito na forma matricial

$$\underset{i=1,\dots,n}{(x_i, y_i)} \longrightarrow \begin{cases} y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_n + \varepsilon_n \end{cases} \Leftrightarrow \underbrace{\begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} 1 & x_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix}}_{n \times 2} \times \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{2 \times 1} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}}_{n \times 1} \Leftrightarrow \boldsymbol{Y = X.\beta + \varepsilon}$$

Soma dos quadrados dos erros $S = \sum_{i=1}^{n} \boldsymbol{\varepsilon_i^2} = \boldsymbol{\varepsilon_i^T \varepsilon_i} = \boldsymbol{(Y - X\beta)^T (Y - X\beta)} = \boldsymbol{Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta}$

O valor mínimo para S é obtido da seguinte maneira:

$$\frac{\partial}{\partial \beta} S = 0 \Leftrightarrow \boldsymbol{-2X^T Y + 2X^T X\hat{\beta} = 0 \Leftrightarrow X^T Y = X^T X\hat{\beta} \Leftrightarrow \hat{\beta} = [X^T X]^{-1} X^T Y}$$

10 **Hipóteses básicas do M.R.L.S.**

- Linearidade entre x e y
- $E[\varepsilon_i|x] = 0$ e $var[\varepsilon_i|x] = \sigma^2 > 0$, (independente de x) ⟶ homocedasticidade condicionada
- $\varepsilon_i \sim N(0, \sigma^2)$ ⟶ normalmente distribuídas (e independentes)

Como estimar σ^2 ?

Conhecendo x_i e y_i, obtêm-se as estimativas de $\hat{\beta}_0$ e $\hat{\beta}_1$ e o valor estimado de $\hat{\sigma}^2$ é dado por:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

$*n-2 \longrightarrow$ como temos estimar os dois parâmetros β_0 e β_1 perdemos dois graus de liberdade na estimação e, portanto, o fator de divisão é $n-2$.

11

Exemplo 2 (cont.) retornam ao exemplo 1:

a) Determinar a reta de regressão linear

b) Interpretar os parâmetros do modelo

c) Calcular os resíduos e estimar a sua variância

Prod (i)	Tam. de lote (x_i)	Horas de trab. (y_i)	$\hat{y}_i$	ε_i	ε_i^2
1	30	73	70	3	9
2	20	50	50	0	0
3	60	128	130	-2	4
4	80	170	170	0	0
5	40	87	90	-3	9
6	50	108	110	-2	4
7	60	135	130	5	25
8	30	69	70	-1	1
9	70	148	150	-2	4
10	60	132	130	2	4
Total	**500**	**1100**	**1100**	**0**	**60**

a) Do exemplo anterior $S_{xx} = 3400$ e $S_{xy} = 6800$. Portanto,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{6800}{3400} = 2$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 110 - 2 \times 50 = 10$$

A reta de regressão é $\hat{y} = 10 + 2x$

b) $\hat{\beta}_1 = 2 \longrightarrow$ o aumento de uma unidade na produção de lotes, aumenta em 2 horas o tempo esperado de trabalho.

c) Variância: $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-2} = \frac{60}{10-2} = 7{,}5$

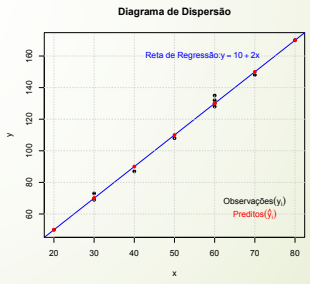
12

Outra resolução (notação matricial)

$$X = \begin{bmatrix} 1 & 30 \\ 1 & 20 \\ 1 & 60 \\ 1 & 80 \\ 1 & 40 \\ 1 & 50 \\ 1 & 60 \\ 1 & 30 \\ 1 & 70 \\ 1 & 60 \end{bmatrix}; \quad Y = \begin{bmatrix} 73 \\ 50 \\ 128 \\ 170 \\ 87 \\ 108 \\ 135 \\ 69 \\ 148 \\ 132 \end{bmatrix}; \quad X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 30 & 20 & 60 & 80 & 40 & 50 & 60 & 30 & 70 & 60 \end{bmatrix} \begin{bmatrix} 1 & 30 \\ 1 & 20 \\ 1 & 60 \\ 1 & 80 \\ 1 & 40 \\ 1 & 50 \\ 1 & 60 \\ 1 & 30 \\ 1 & 70 \\ 1 & 60 \end{bmatrix} = \begin{bmatrix} 10 & 500 \\ 500 & 28400 \end{bmatrix}$$

$$\left.\begin{array}{r} [X^T X]^{-1} = \begin{bmatrix} 0{,}84 & -0{,}01 \\ -0{,}01 & 0{,}00 \end{bmatrix} \\ X^T Y = \begin{bmatrix} 1100 \\ 61800 \end{bmatrix} \end{array}\right\} \Rightarrow \hat{\beta} = [X^T X]^{-1} X^T Y = \begin{bmatrix} 10 \\ 2 \end{bmatrix}$$

Assim, a equação da regressão é $\hat{y} = 10 + 2x$



13

b) Inferência sobre os parâmetros

Para realizar inferência sobre os parâmetros, é necessário conhecer a distribuição amostral dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$. Considere o modelo de regressão linear $\boldsymbol{Y} = \boldsymbol{X}.\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, cujo estimadores de $\boldsymbol{\beta}$ são dados por $\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

- Valor esperado

$$E(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{Y} = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} + [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon} = \boldsymbol{\beta} + [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\overbrace{E(\boldsymbol{\varepsilon}|\boldsymbol{X})}^{=0} = \boldsymbol{\beta}$$

- Variância/covariância

$$\Sigma_{\widehat{\boldsymbol{\beta}}} = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{Y} = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\overbrace{var(\boldsymbol{\varepsilon}|\boldsymbol{X})}^{=\sigma^2}[\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X} = \overbrace{[\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{X}}^{I}[\boldsymbol{X}^T\boldsymbol{X}]^{-1}\sigma^2 = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\sigma^2$$

Assim, a matriz covariância dos parâmetros, é dada por: $\Sigma_{\widehat{\boldsymbol{\beta}}} = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\sigma^2 = \begin{bmatrix} var(\beta_0) & cov(\beta_0,\beta_1) \\ cov(\beta_0,\beta_1) & var(\beta_1) \end{bmatrix}$

$$[\boldsymbol{X}^T\boldsymbol{X}] = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix};\ [\boldsymbol{X}^T\boldsymbol{X}]^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}}{s_{xx}} & \frac{-\bar{x}}{s_{xx}} \\ \frac{-\bar{x}}{s_{xx}} & \frac{1}{s_{xx}} \end{bmatrix}$$

c_{ii} ⟶ elementos da diagonal principal da $\Sigma_{\widehat{\boldsymbol{\beta}}}$

c_{ij} ⟶ elementos que não sejam da diagonal principal

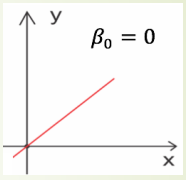
$var(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2 c_{ii}$ e $cov(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2 c_{ij};\ i \neq j.$

Logo, a distribuição de cada parâmetro $\hat{\beta}_0$ e $\hat{\beta}_1$ é $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} \sim N(\boldsymbol{\beta}_i, \sigma^2 c_{i+1,i+1}), i = 0{,}1.$

14

Teste de hipóteses e intervalos de confiança para os parâmetros de regressão

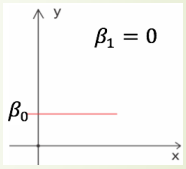
Será que os coeficientes de regressão são estatisticamente significativos ?



- **Teste de hipóteses, significância α**

Hipóteses a testar:

$H_0: \beta_j = 0$ vs. $H_1: \beta_j \neq 0; j = 0, 1$



Estatística de teste:

$\hat{\beta}_j \sim N(\beta_j, \sigma^2_{\hat{\beta}_j}) \Rightarrow T_j = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}} \sim t_{(n-2)}$, onde $S_{\hat{\beta}_j}$ é desvio padrão de $\hat{\beta}_j$

$$S_{\hat{\beta}_j} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \text{ se } j = 0 \text{ e } S_{\hat{\beta}_j} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \text{ se } j = 1$$

Decisão: Rejeitar H_0 em favor de H_1 se $|T_{jobs}| > t_{(1-\alpha/2,n-2)}$. Caso contrário não se rejeitar H_0.

- **Intervalos de confiança $(1-\alpha)$**

$$\left[\hat{\beta}_j - t_{(1-\alpha/2,n-2)}S_{\hat{\beta}_j}\,;\hat{\beta}_j + t_{(1-\alpha/2,n-2)}S_{\hat{\beta}_j}\right]$$

15

Exemplo 3 (cont.) - retornam ao exemplo 1, testar se:

a) o intercepto é estatisticamente significativo

b) o declive é estatisticamente significativo

Prod. (i)	Tam. de lote (x_i)	Horas de trab. (y_i)	x_i^2	$x_i y_i$	$\hat{y}_i$	ε_i	ε_i^2
1	30	73	900	2190	70	3	9
2	20	50	400	1000	50	0	0
3	60	128	3600	7680	130	-2	4
4	80	170	6400	13600	170	0	0
5	40	87	1600	3480	90	-3	9
6	50	108	2500	5400	110	-2	4
7	60	135	3600	8100	130	5	25
8	30	69	900	2070	70	-1	1
9	70	148	4900	10360	150	-2	4
10	60	132	3600	7920	130	2	4
Total	**500**	**1100**	**28400**	**61800**	**1100**	**0**	**60**

Dos exemplos anteriores:

- $\hat{\beta}_0 = 10$
- $\hat{\beta}_1 = 2$
- $\hat{\sigma}^2 = 7{,}5$
- $\bar{x}^2 = 2500$
- $S_{xx} = 3400$

Portanto,

$$S_{\hat{\beta}_0} = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} = 2{,}50$$

$$S_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = 0{,}047$$

16

a) O intercepto β_0 é estatisticamente significativo?

Hipóteses a testar:

$H_0: \beta_0 = 0$ vs. $H_1: \beta_0 \neq 0$

A estatística de teste:

$$T_{0obs} = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} = \frac{10}{2{,}50} \approx 4 \longrightarrow |4| = 4$$

Valor crítico:

$$\alpha = 0{,}05 \Longrightarrow t_{(1-\alpha/2, n-2)} = t_{(0{,}975\,,\,8)} \approx 2{,}306$$

Decisão: como $|T_{1obs}| > t_{(0{,}975\,,\,8)}$, então rejeitar-se H_0 em favor de H_1 com 5% significância. O intercepto é significativo no modelo de regressão.

I.C.: $[\hat{\beta}_0 - t_{(1-\alpha/2, n-2)} S_{\hat{\beta}_0}\,;\, \hat{\beta}_0 + t_{(1-\alpha/2, n-2)} S_{\hat{\beta}_0}]$ = [4,23; 15,77]

b) O declive β_1 é estatisticamente significativo?

Hipóteses a testar:

$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

A estatística de teste:

$$T_{1obs} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{2}{0{,}047} \approx 42{,}58 \longrightarrow |42{,}58| = 42{,}58$$

Decisão: como $|T_{1obs}| > t_{(0{,}975\,,\,8)}$, então rejeitar-se H_0 em favor de H_1 com 5% significância. O declive é significativo no modelo de regressão.

I.C.: $[\hat{\beta}_1 - t_{(1-\alpha/2, n-2)} S_{\hat{\beta}_1}\,;\, \hat{\beta}_1 + t_{(1-\alpha/2, n-2)} S_{\hat{\beta}_1}]$ = [1,89; 2,11]

17

3.3. Significado, avaliação da qualidade e validação dos pressupostos do modelo

1) Significado estatístico do modelo
   - Teste ao declive β_1
   - Teste ANOVA da regressão
2) Avaliação da qualidade do modelo
   - Coeficiente de determinação
   - Coeficiente de determinação ajustado
3) Validação dos pressupostos do modelo – análise de resíduos
   - $E[\varepsilon_i|x] = 0$ e $var[\varepsilon_i|x] = \sigma^2, i = 1, \dots, n$
   - $\varepsilon_i \sim N(0, \sigma^2)$
   - ε_i são independentes

18

a) Significado estatístico do modelo

- **Teste ao declive $\boldsymbol{\beta_1}$**

Verificar se as variáveis independentes x_i contribuem significativamente com informação para explicar linearmente a variação da variável resposta y_i, como o que já referido anteriormente.

Hipóteses a testar

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

Estatística de teste:

$$T_1 = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \sim t_{(n-2)}$$

Decisão: com α significância rejeitar H_0 em favor de H_1 se $|T_{1obs}| > t_{1-\alpha/2, n-2}$. Caso contrário não se rejeita H_0 em favor de H_1.

19

- **Teste ANOVA da regressão**

Avaliar a significância estatística do modelo de regressão

Hipóteses a testar

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

Estatística de teste:

$$F = \frac{MQ_R}{MQ_E} \sim F_{1,n-2}$$

Decisão: Rejeitar H_0 em favor de H_1 se $F_{obs} > F_{1-\alpha,1,n-2}$. Caso contrário não se rejeita H_0.

Tabela ANOVA

Fonte de variação	SQ	$g.l.$	MQ	F_{obs}	$valor - P$
Regressão (explicada)	SQ_R	1	$MQ_R = \frac{SQ_R}{1}$	$\frac{MQ_R}{MQ_E}$	$P(F > F_{obs})$
Erros (não explicada)	SQ_E	$n-2$	$MQ_E = \frac{SQ_E}{n-2}$		
Total	SQ_T	$n-1$			

Nota: $SQ_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$; $SQ_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2$; $SQ_T = SQ_R + SQ_E = S_{yy}$; $MQ_E = \hat{\sigma}^2$

20

Exemplo 4 (cont.) – retornam ao exemplo 1

Avaliar a significância estatística do modelo de regressão.

Prod. (i)	**Tam. lote (x_i)**	**Horas de trab. (y_i)**	$\hat{y}_i$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
1	30	73	70	-40	1600
2	20	50	50	-60	3600
3	60	128	130	20	400
4	80	170	170	60	3600
5	40	87	90	-20	400
6	50	108	110	0	0
7	60	135	130	20	400
8	30	69	70	-40	1600
9	70	148	150	40	1600
10	60	132	130	20	400
Total	**500**	**1100**	**1100**		**13600**

$\bar{y} = 110$; $\hat{\sigma}^2 = 7{,}5$ e $SQ_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 13600$

Hipóteses a testar:

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

A estatística de teste:

$$F_{obs} = \frac{SQ_R}{\hat{\sigma}^2} = \frac{13600}{7{,}5} \approx 1813{,}33$$

Valor crítico:

$$\alpha = 0{,}05 \Longrightarrow F_{1-\alpha,1,n-2} = F_{0.95,1,8} \approx 5{,}32$$

Decisão: como $F_{obs} > F_{0.95,1,8}$, concluímos rejeitar H_0 em favor de H_1, com 5% de significância. Logo, o modelo é estatisticamente significativo.

21

b) Avaliação da qualidade do modelo

- **Coeficiente de determinação** ⟶ Quantificar percentagem da variação de y que é explicada pelo modelo.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SQ_R}{S_{yy}}$$

Também pode ser escrito como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SQ_E}{S_{yy}}$$

- Limites de variação: $0 \leq R^2 \leq 1$.
- $R^2 \cong 1$, a maior parte da variação de y *é explicada linearmente* por x.
- $R^2 \cong 0$, a maior parte da variação de y *não é explicada linearmente* por x.

22

- **Coeficiente de determinação ajustado** ⟶ Quantificar o grau do ajustamento do modelo. É definido a partir de R^2 e ajustado com base na dimensão da amostra (n) e no número de parâmetros do modelo (p). No caso do modelo de regressão simples, $p = 2$.

$$R_a^2 = 1 - \frac{SQ_E/(n-p)}{S_{yy}/(n-1)}$$

- Limites de variação: $0 \leq R_a^2 \leq 1$.
- $R_a^2 \cong 1$, o modelo é bastante adequado.
- $R_a^2 \cong 0$, o modelo é pouco adequado.

23

Exemplo 5 (cont.) – retornam ao exemplo 1

a) Quantificar a variação de y explicada pelo modelo.

b) Quantificar o grau de ajustamento do modelo aos dados.

Prod (i)	**Tam lote (x_i)**	**Hor. trab. (y_i)**	$\hat{y}_i$	ε_i^2
1	30	73	70	9
2	20	50	50	0
3	60	128	130	4
4	80	170	170	0
5	40	87	90	9
6	50	108	110	4
7	60	135	130	25
8	30	69	70	1
9	70	148	150	4
10	60	132	130	4
Tot.	**500**	**1100**	**1100**	**60**

a) Dos exemplos anteriores, $SQ_R = 13600$ e $S_{yy} = 13660$. Portanto,

$$R^2 = \frac{SQ_R}{S_{yy}} = \frac{13600}{13660} \approx 0{,}996 \text{ ou } R^2 = 1 - \frac{SQ_E}{S_{yy}} = 1 - \frac{60}{13660} \approx 0{,}996$$

Assim, concluímos que cerca de 99,6% da variabilidade de y é explicada pelo modelo.

b) Grau de ajustamento: $R_a^2 = 1 - \frac{SQ_E/(n-2)}{S_{yy}/(n-1)} = 1 - \frac{\frac{60}{8}}{\frac{13660}{9}} \approx 0{,}995$

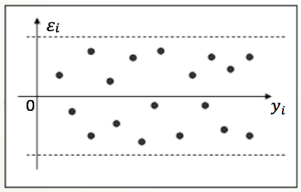
O grau de ajustamento é aproximadamente $0{,}995$. Logo, o modelo é bastante adequado.

24

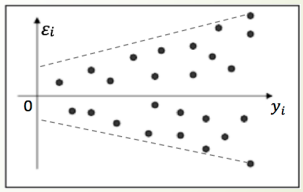
c) Validação dos pressupostos do modelo - análise de resíduos

- $\boldsymbol{E[\varepsilon_i|x] = 0}$ **e** $\boldsymbol{var[\varepsilon_i|x] = \sigma^2, i = 1, \ldots, n}$ ⟶ Gráficamente

Os pontos do gráfico devem distribuir-se de forma aleatória em torno da reta que corresponde ao resíduo zero, formando uma mancha de largura uniforme. Dessa forma, será de esperar que os erros sejam independentes, de média nula e de variância constante.



Homocedasticidade
(variância constante)



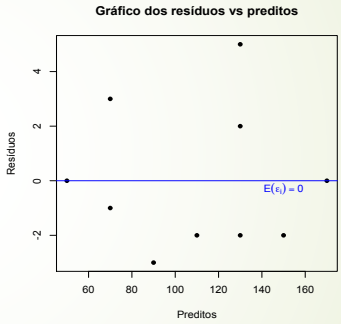
Heterocedasticidade
(variância não constante)

25

Exemplo 6 (cont.) – retornam ao exemplo 1

Verificar se os resíduos são independentes, têm média zero e variância constante.

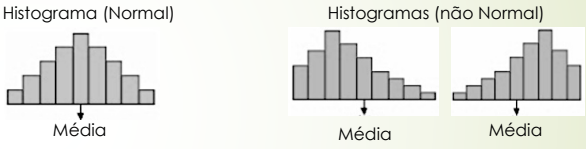
Prod. (i)	Tam. lote (x_i)	Horas de trab. (y_i)	$\hat{y}_i$	ε_i
1	30	73	70	-3
2	20	50	50	0
3	60	128	130	-2
4	80	170	170	0
5	40	87	90	-3
6	50	108	110	-2
7	60	135	130	5
8	30	69	70	-1
9	70	148	150	-2
10	60	132	130	2
Total	**500**	**1100**	**1100**	**0**



Da inspeção visual do gráfico dos resíduos, não há razão para duvidar que os pressupostos não são válidos.

26

- $\varepsilon_i \sim N(0, \sigma^2)$ ⟶ Graficamente
  - ✓ No **histograma** dos ε_i, se os erros possuírem distribuição Normal: observa-se - Concentração de valores em torno de um valor central; - Simétrica em torno do valor central; - Frequência pequena de valores muito extremos.



  - ✓ No **normal QQ-plot**, se os erros possuírem distribuição Normal, todos os pontos do gráfico devem posicionarem-se sobre a reta de referência.

27

- $\boldsymbol{\varepsilon_i \sim N(0, \sigma^2)}$ ⟶ Teste estatístico

H_0 : a amostra provem de uma distribuição normal

H_1: $\sim H_0$

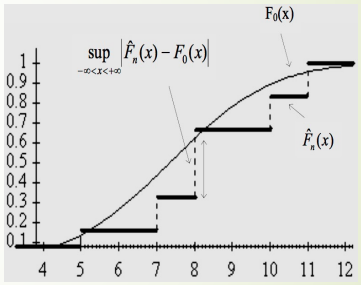
Teste de hipóteses:

- Teste Kolmogorov-Smirnov (KS) ⟶ teste construído para qualquer distribuição amostras grandes.
- Teste de normalidade de Lilliefors ⟶ estatística do teste KS adaptada para o teste normalidade (teste adequado para $n \geq 30$)
- Teste de normalidade de Shapiro-Wilk ⟶(teste adequado para $n < 30$)

28

Teste de Kolmogorov-Smirnov (KS)

Consiste na comparação de função de distribuição acumulada (f.d.a.) dos valores observados e função de distribuição acumulada teórica, de acordo com H_0, e na determinação do ponto de maior distância vertical entre as duas funções.



Hipóteses a testar:

$$H_0 : F(x) = F_0(x) \text{ vs. } H_1 : F(x) \neq F_0(x)$$

Estatística de teste:

$$D_n = \sup_x |\hat{F}_n(x) - F_0(x)|$$

Onde: $\hat{F}_n(x) = \frac{\#\, observações \leq x}{n}$ é f.d.a. empírica

$F_0(x) = \mathrm{P}(\mathrm{X} \leq x)$ é f.d.a. em H_0

Decisão: rejeita-se H_0 em favor de H_1 se "D_n for um valor muito alto", concretamente se $D_n \geq d_{\alpha,n}$ (tabelado). Caso contrário não se rejeita H_0.

29 **Teste de normalidade de Lilliefors ($n \geq 30$)**

O teste de Lilliefors usa a mesma estatística do Teste de KS, mas a tabela de valores críticos é a Tabela de Teste de Lilliefors, que é usada em vez da Tabela de KS.

Teste de normalidade de Shapiro-Wilk ($n < 30$)

É uma alternativa ao teste de KS para testar se a variável em estudo na amostra aleatória possui, ou não, distribuição normal e é adequado para amostras pequenas.

Hipóteses a testar:

$H_0 : X \sim N(\mu, \sigma^2)$ vs. $H_1 : \sim H_0$

Estatística de teste:

$W = \frac{b^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, onde $b = \begin{cases} \sum_{i=1}^{\frac{n}{2}} a_{n-i+1}\left(x_{(n-i+1)} - x_{(i)}\right) & se\ n\ é\ par \\ \sum_{i=1}^{\frac{n+1}{2}} a_{n-i+1}\left(x_{(n-i+1)} - x_{(i)}\right) & se\ n\ é\ impar \end{cases}$; em que a_{n-i+1} são tabelados

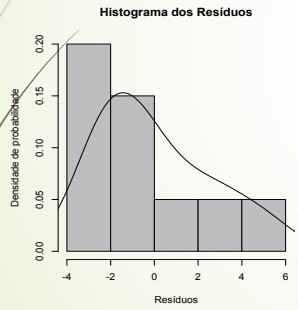
x_i estão ordenados por ordem crescente

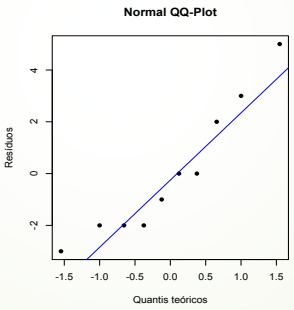
Decisão: rejeitar H_0 em favor de H_1 se $W_{obs} < W_{\alpha,n}$ (tabelado), caso contrário não se rejeita H_0.

30 **Exemplo 7 (cont.) – retornam ao exemplo 1**

Será que os resíduos possuem uma distribuição Normal?

- Observar através do histograma ou QQ-plot





Da inspeção visual do histograma não é evidente que os resíduos provêm de uma distribuição normal, mas existe uma proximidade dos pontos em relação à linha de referência do QQ-plot.

31

➢ Teste de Normalidade

Como $n = 10 < 30$, logo aplica-se o teste de normalidade de *Shapiro-Wilk*.

Hipóteses a testar:

$H_0 : \boldsymbol{\varepsilon} \sim N(0\,, 7.5)$ vs. $H_1 : \sim H_0$

Estatística de teste:

$$W_{obs} = \frac{b^2}{\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2} = \frac{b^2}{\boldsymbol{\varepsilon_i^2}}$$

$$= \frac{54{,}4319}{60} \approx 0{,}9072$$

Valor crítico:

$$\alpha = 0{,}0 \Longrightarrow W_{\alpha,n} = W_{0.05,10} = 0{,}842$$

$n = 10$ é par

$$\Longrightarrow b = \sum_{i=1}^{\frac{n}{2}} a_{n-i+1}\left(\varepsilon_{(n-i+1)} - \varepsilon_{(i)}\right)$$

$$\Leftrightarrow b = \sum_{i=1}^{5} a_{10-i+1}\left(\varepsilon_{(10-i+1)} - \varepsilon_{(i)}\right)$$

$$\Leftrightarrow b = a_{10-1+1}(\varepsilon_{10} - \varepsilon_1) + a_9(\varepsilon_9 - \varepsilon_2) + \cdots + a_{10-5+1}(\varepsilon_6 - \varepsilon_5)$$

$$= 0{,}5739(5 + 3) + \cdots + 0{,}0399(-1 + 2) = 7{,}3778$$

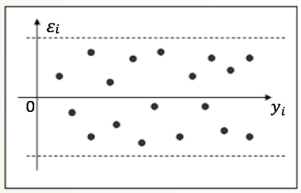
Decisão: Como $W_{obs} > W_{0.05,10}$, então não se rejeita H_0 em favor de H_1, com 5% de significância. Logo, assume-se que os resíduos possuem uma distribuição normal e o pressuposto fica validado.

32

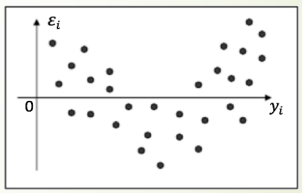
- **Independência de ε_i ⟶ Graficamente**

Se os resíduos forem independentes, os pontos do diagrama de dispersão $\boldsymbol{\varepsilon_i}$ versus $\boldsymbol{y_i}$ têm uma comportamento aleatório, i.e., não seguem nenhum padrão na sua distribuição em torno da linha central.

Têm comportamento aleatório



Não têm comportamento aleatório



33

- **Independência de ε_i → teste estatístico**

✓ **Teste de Durbin-Watson:** se os resíduos forem independentes, a magnitude de um resíduo não influencia a magnitude do resíduo seguinte. Neste caso, a correlação entre resíduos sucessivos é nula (ρ = 0).

Hipóteses a testar:

$H_0 : \rho = 0$

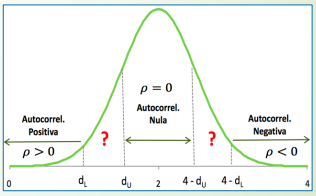
$H_1 : \rho \neq 0$

Estatística de teste: $dw = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2}$ onde dw é um valor tal que $0 \leq dw < 4$

Valores críticos: d_L e d_U (tabelados)

Decisão:

- $d_U \leq dw < 4 - d_U$ → não se rejeita H_0
- $0 \leq dw < d_L$ ou $4 - d_L \leq dw < 4$ → rejeita-se H_0
- $d_L \leq dw < d_U$ ou $4 - d_U \leq dw < 4 - d_L$ → nada se pode concluir



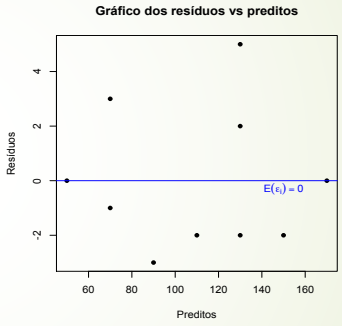
34

Exemplo 8 (cont.) – retornam ao exemplo 1

Será que ao resíduos são independentes ?

Prod. (i)	Tam. lote (x_i)	Horas de trab. (y_i)	$\hat{y}_i$	ε_i
1	30	73	70	-3
2	20	50	50	0
3	60	128	130	-2
4	80	170	170	0
5	40	87	90	-3
6	50	108	110	-2
7	60	135	130	5
8	30	69	70	-1
9	70	148	150	-2
10	60	132	130	2
Total	**500**	**1100**	**1100**	**0**

➢ Gráfico dos resíduos



Da figura, observa-se que a distribuição dos resíduos não segue nenhum padrão, estando distribuídos de forma aleatória. Isto indica que o pressuposto de independência não foi violado.

35

➢ Aplicação do teste de Durbin-Watson

ε_i	ε_i^2	$\varepsilon_i - \varepsilon_{i-1}$	$(\varepsilon_i - \varepsilon_{i-1})^2$
3	9	-	-
0	0	-3	9
-2	4	-2	4
0	0	2	4
-3	9	-3	9
-2	4	1	1
5	25	7	49
-1	1	-6	36
-2	4	3	9
2	4	0	0
Total	**60**		**129**

Hipóteses a testar:

H_0 : existe independência vs. H_1 : $\sim H_0$

Estatística de teste:

$$dw = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2} = \frac{129}{60} = 2{,}15$$

Valor crítico:

$\alpha = 0{,}05; n = 10; k = 1$ (# de variáveis regressoras $) \Longrightarrow d_L = 0{,}88; d_U = 1{,}32$

Como $d_U \leq dw < 4 - d_U \Leftrightarrow 1{,}32 < 2{,}15 < 2{,}68$, não se rejeita H_0 em favor de H_1. Assim, não há evidência de que o pressuposto da independência dos resíduos não seja válido.

36

3.4. Previsão na média e previsão pontual

Ao assumir que os erros tem distribuição Normal concluímos que também as observações Y_i vão ter distribuição Normal já que $\boldsymbol{Y} = \boldsymbol{X}.\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, e a soma de uma Normal com uma constante tem distribuição Normal.

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} \sim N\left(\boldsymbol{\beta}_i, \sigma^2 c_{i+1,i+1}\right), i = 0{,}1.$$
$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2)$$

Distribuição amostral de y_i é dada por:

$$\left.\begin{array}{l} E(\boldsymbol{Y}|\boldsymbol{X}) = E(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}|\boldsymbol{X}) = E(\boldsymbol{X\beta}) + \boldsymbol{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{X\beta} \\ var(\boldsymbol{Y}|\boldsymbol{X}) = var(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}|\boldsymbol{X}) = \overbrace{var(\boldsymbol{X\beta}|\boldsymbol{X})}^{\substack{(constante) \\ =0}} + var(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \overbrace{var(\boldsymbol{\varepsilon}|\boldsymbol{X})}^{=\sigma^2} = \sigma^2 \end{array}\right\} \Longrightarrow \boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2)$$

37

a) Previsão em média (previsão do valor esperado)

No caso previsão em média, pretende-se estimar o parâmetro
$\theta = E(y_i|x_{i1} = c_1, x_{i2} = c_2, \dots, x_{ik} = c_k) = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k$,
com k é número de regressores.

No caso de modelo regressão linear simples $k = 1$. Neste caso, o estimador de θ é

$$\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_1 c_1.$$

Fazendo $c = [1 \quad c_1]$ então o valor esperado e variância do estimador de θ são seguintes:

$$E(\hat{\theta}|x, c) = E(\hat{\beta}_0 + \hat{\beta}_1 c_1|x, c) = \beta_0 + \beta_1 c_1 = \theta$$

$$Var(\hat{\theta}|x, c) = Var(c\hat{\beta}|x, c) = c\, Cov(\hat{\beta}|x, c)c^T = \sigma^2 c(X^T X)^{-1} c^T \dots\dots\dots\dots (*)$$

A raiz quadrada de (*) é o **erro padrão da previsão em média,**

$$s_{\hat{\theta}} = \hat{\sigma}\sqrt{c(X^T X)^{-1} c^T}$$

e, assim, o I.C. de previsão para θ com confiança $(1 - \alpha)$ é dado por

$$\left[\hat{\theta} - t_{1-\alpha/2, s_{\hat{\theta}}}\,; \hat{\theta} + t_{1-\alpha/2, n-2} s_{\hat{\theta}}\right].$$

38

b) Previsão pontual (previsão individual)

Considere-se de novo que $x_{i1} = c_1, x_{i2} = c_2, \dots, x_{ik} = c_k$, ou seja, neste caso $k = 1$, então $y_0 = \beta_0 + \beta_1 c_1 + \varepsilon_0$
Enquanto que na previsão em média se pretendia estimar $E(y_0|x, c)$, na previsão pontual procura-se prever os valores assumidos por y_0.
Considere-se o previsor mínimo quadrado de y_0,

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1$$

e o erro de previsão,

$$\varepsilon = y_0 - \hat{y}_0$$

utilizando a variável aleatória ε, vão estudar-se as propriedades estatísticas do previsor. Como

$$E(\varepsilon|x, c) = E(y_0 - \hat{y}_0|x, c) = 0$$

a variância de ε, condicionada por X e c, é dada por

$$Var(\varepsilon|x, c) = \sigma^2 \left\{1 + c(X^T X)^{-1} c^T\right\}$$

Assim, o **erro padrão da previsão** é dado por

$$s_\varepsilon = \hat{\sigma}\sqrt{1 + c(X^T X)^{-1} c^T}$$

Portanto, o I.C. de previsão para y_0 com confiança $(1 - \alpha)$ é dado por

$$\left[\hat{y}_0 - t_{1-\alpha/2, s_\varepsilon}\,; \hat{y}_0 + t_{1-\alpha/2, n-2} s_\varepsilon\right].$$

39

Exemplo 9 (cont.) – retornam ao exemplo 1

Calcule o número de horas de trabalho previsto para a produção de um lote de tamanho 40. comenta os resultados.

Resoluções:

$x_i = 40 \longrightarrow c = [1 \;\; 40];$

$n = 10;\ \alpha = 0{,}05$

Modelo:

$\hat{y} = 10 + 2x = 10 + 2{\times}40 = 90 = \hat{y}_0$

$$[X^TX]^{-1} = \begin{bmatrix} 0{,}83529 & -0{,}01471 \\ -0{,}01471 & 0{,}00029 \end{bmatrix}$$

$c[X^TX]^{-1}c^T = 0{,}129$

$\hat{\sigma} = 2{,}739$

$t_{1-\alpha/2,n-2} = t_{0{,}975{,}8} = 2{,}306$

- Previsão em média (95% confiança)

  $s_{\hat{y}_0} = \hat{\sigma}\sqrt{c(X^TX)^{-1}c^T} = 2{,}739{\times}\sqrt{0{,}129} = 0{,}985$

  $\therefore \text{I.C.}: \left[\hat{y}_0 \pm t_{1-\alpha/2,n-2}s_{\hat{y}_0}\right] = [87{,}729\,;92{,}272\,]$

  $p[87{,}729 \leq \hat{y}_0 \leq 92{,}272\,] = 0{,}95$

- Previsão pontual

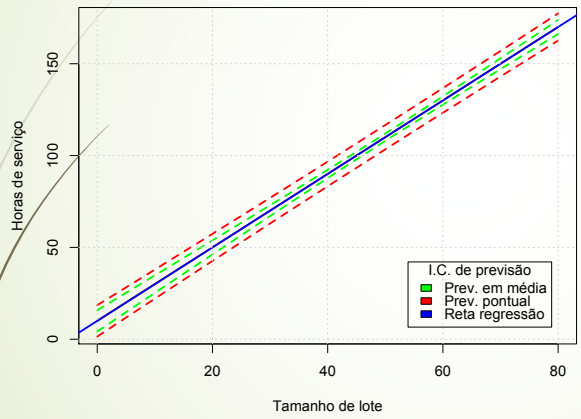
  $s_{\hat{y}} = \hat{\sigma}\sqrt{1 + c(X^TX)^{-1}c^T} = 2{,}739{\times}\sqrt{1 + 0{,}129} = 2{,}910$

  $\therefore \text{I.C.}: \left[\hat{y} \pm t_{1-\alpha/2,n-2}s_{\hat{y}}\right] = [83{,}289\,;96{,}711\,]$

  A probabilidade do intervalo [87,729 ; 92,272] conter o valor de previsão (em média) para produção de um lote de tamanho 40 é 95%. Para a previsão pontual a interpretação é equivalente.

40

Representação dos intervalos de confiança 95% de previsão



Note que os resultados obtidos permitem mostrar que a amplitude do I.C. de previsão pontual é sempre maior do que o respetivo I.C. da previsão em média, uma vez que $\hat{\sigma}\sqrt{1 + c(X^TX)^{-1}c^T} > \hat{\sigma}\sqrt{c(X^TX)^{-1}c^T}$.

41 **Exercício 1**

A massa muscular dos adultos decresce, em geral, com a idade. Com o objetivo de averiguar tal relação em mulheres, um nutricionista selecionou aleatoriamente 16 mulheres com idades entre os 40 e os 79 anos. Os resultados observados constam da seguinte tabela:

Idade (x)	71	64	43	67	56	73	68	56	76	65	45	58	45	53	49	78
Medida M.M.(y)	82	91	100	68	87	73	78	80	65	84	116	76	97	100	105	77

1. Verificar se existe relação linear entre massa muscular e idade, quantificar o grau de relacionamento e comentar o resultado obtido.
2. No caso de existir correlação:
   a) Determinar os parâmetros do modelo de regressão linear e interpretar o coeficiente de regressão.
   b) Calcular os erros padrão dos parâmetros do modelo.
   c) Testar se os parâmetros do modelo são estatisticamente significativos, indicando também os respetivos intervalos de confiança a 95%.

42

d) Testar se o modelo é estatisticamente significativo.
e) Quantificar a percentagem da variação explicada pelo modelo.
f) Quantificar o grau de ajustamento do modelo aos dados.
g) Determinar os resíduos, e averiguar se os resíduos:
   i. têm média zero e variância constante,
   ii. São normalmente distribuídos,
   iii. são independentes.
h) Determinar as previsões de medida de massa de muscular para mulheres que têm 68 anos de idade e comentor os resultados obtidos.

Soluções:

1. $r = -0{,}82$
2. a) $\beta_0 = 148{,}0507$; $\beta_1 = -1{,}0236$
   b) $s_{\beta_0} = 11{,}5629$; $s_{\beta_1} = 0{,}1882$
   c) t_{0obs} =12,804; I.C.: [123,2507 ; 172,8507]
   $t_{1obs} = -5{,}439$; I.C.: [-1,4272 ; - 0,6200]

d) $F_{0obs} = 29{,}59$
e) 67,88%
f) 65,59%
g) ii. W = 0,9648
   iii. Dw = 1,5547
h) Prev. Em média, I.C.: [73,0307 ; 83,8625]
   Prev. pontual, I.C.: [59,7494 ; 97,1438]

43

Exercício 2

Pretende-se construir um modelo explicativo do custo de manutenção de rebocadores de navios em função da idade do rebocador. Os resultados da estimação com o *software R* são os seguintes:

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-35.604 -20.785  -6.169  13.208  53.920

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.397     14.680   2.207  0.04333 *
x             13.177      3.557   3.704  0.00212 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.32 on 15 degrees of freedom
Multiple R-squared:  0.4777,   Adjusted R-squared:  0.4429
F-statistic: 13.72 on 1 and 15 DF,  p-value: 0.002122
```

1) Escreva o modelo na forma: $y = \beta_0 + \beta_1 x + \varepsilon$
2) Diga o significado do valor do parâmetro β_1.
3) Considerando nível de significância 5%, os parâmetros do modelo são estatisticamente significativos ?
4) Indique os erros padrão das estimativas dos parâmetros.
5) Qual a percentagem de variabilidade das observações que é explicada pelo modelo linear.
6) Em nível de significância 5%, será que o modelo é estatisticamente significativo?

44

4. Regressão linear Múltipla

4.1. Modelo de Regressão linear Múltipla (M.R.L.M.)

Um M.R.L.M. é um modelo linear com mais do que um regressor. Neste caso, relaciona-se uma variável independente y com mais do que uma variável dependente $x_j, j = 1, \dots, k$.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

- $i = 1, \dots, n$ é o número de observações
- $j = 1, \dots, k$ é o numero de regressores
- x_{ij} é o valor de observação i na variável x.
- β_j são os parâmetros da regressão
- $p = k + 1$ é o número de parâmetros do modelo

Em notação matricial, o M.R.L.M. escrever-se de forma análogo ao M.R.L.S.

$$\begin{matrix}(x_{ij}, y_i)\\ i = 1, \dots, n\\ j = 1, .., k\end{matrix} \longrightarrow \begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \cdots \beta_k x_{1k} + \varepsilon_1 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_k x_{nk} + \varepsilon_n \end{cases} \Leftrightarrow \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{nk} & \dots & x_{nk} \end{bmatrix}}_{n \times p} \times \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{p \times 1} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{n \times 1} \Leftrightarrow \boldsymbol{Y = X.\beta + \varepsilon}$$

45

4.2. Estimação e inferência sobre os parâmetro

- As estimativas de $\beta_0,\ \beta_1, \dots, \beta_k$ que minimizam a soma dos quadrados dos erros são

$$\hat{\boldsymbol{\beta}} = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

e, assim, a reta da regressão é dada por $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$

- As somas dos quadrados em forma matricial são dados por

$$SQ_T = \boldsymbol{Y}^T\boldsymbol{Y} - n\bar{y}^2\ ;\ SQ_R = \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} - n\bar{y}^2\ ;\ \ SQ_E = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y}$$

onde $\hat{\sigma}^2 = \frac{SQ_E}{n-p} = MQ_E$

- Distribuição amostral de $\hat{\beta}_j$

$\hat{\boldsymbol{\beta}}_{\boldsymbol{j}} \sim N(\boldsymbol{\beta}_j, \sigma^2 c_{j+1,j+1}); j = 0,1, \dots, k$. Onde $c_{j+1,j+1}$ é elemento da diagonal principal da

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2[\boldsymbol{X}^T\boldsymbol{X}]^{-1} \text{ é } \begin{bmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0,\hat{\beta}_1) & \cdots & cov(\hat{\beta}_0,\hat{\beta}_k) \\ cov(\hat{\beta}_1,\hat{\beta}_0) & var(\hat{\beta}_1) & \cdots & cov(\hat{\beta}_1,\hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{\beta}_k,\hat{\beta}_0) & cov(\hat{\beta}_k,\hat{\beta}_1) & \cdots & var(\hat{\beta}_k) \end{bmatrix}.$$

46

Teste de hipóteses e intervalos de confiança para os parâmetros da regressão

- **Teste de hipóteses para $\boldsymbol{\beta_j}$, significância $\boldsymbol{\alpha}$**

Hipóteses a testar:

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0$

Estatística de teste: $\hat{\beta}_j \sim N\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right) \Rightarrow t = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \sim t_{(n-p)}, j = 0,1, \dots k$, onde $S_{\hat{\beta}_j} = \hat{\sigma}\sqrt{c_{j+1,j+1}}$

Decisão: Rejeitar H_0 em favor de H_1 se $|t_{obs}| > t_{(1-\alpha/2, n-p)}$. Caso contrário não rejeitar H_0

- **Intervalos de confiança $(\boldsymbol{1-\alpha})$ para $\boldsymbol{\beta_j}$**

$$\left[\hat{\beta}_j - t_{(1-\alpha/2, n-p)} S_{\hat{\beta}_j}\ ;\ \hat{\beta}_j + t_{(1-\alpha/2, n-p)} S_{\hat{\beta}_j}\right]$$

47

4.3. Significado e avaliação da qualidade da regressão

a) Significado estatístico do modelo

Teste ANOVA da regressão – equivalente ao referido no M.R.L.S.

Hipóteses a testar

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

$H_1 : \beta_j \neq 0$ para algum j

Estatística de teste: $F = \frac{MQ_R}{MQ_E} \sim F_{k,n-p}$

Decisão: Rejeitar H_0 em favor de H_1 se $F_{obs} > F_{1-\alpha,k,\,n-p}$. Caso contrário não rejeitar H_0

Tabela ANOVA

Fonte de variação	SQ	$g.l.$	MQ	F_{obs}	$P-value$
Regressão (explicada)	SQ_R	k	$MQ_R = \frac{SQ_R}{k}$	$\frac{MQ_R}{MQ_E}$	$P(F > F_{obs})$
Erros (não explicada)	SQ_E	$n-p$	$MQ_E = \frac{SQ_E}{n-p}$		
Total	SQ_T	$n-1$			

48

b) Avaliação da qualidade do modelo

- **Coeficiente de determinação**

$$R^2 = \frac{SQ_R}{SQ_T}$$

  - Limites de variação: $0 \leq R^2 \leq 1$.
  - $R^2 \cong 1$, significa maior parte da variação de y é *explicada linearmente* por x.
  - $R^2 \cong 0$, significa maior parte da variação de y *não é explicada linearmente* por x.

- **Coeficiente de determinação ajustado**

$$R_a^2 = 1 - \frac{SQ_E/(n-p)}{SQ_T/(n-1)}$$

  - Limites de variação: $0 \leq R_a^2 \leq 1$.
  - $R_a^2 \cong 1$, o modelo é bastante adequado.
  - $R_a^2 \cong 0$, o modelo pouco adequado.

49

4.4. Previsão na média e previsão pontual

a) Previsão na média

Tal como no modelo da regressão linear simples, pretende-se estimar o parâmetro

$$\theta = E(y_i|x_{i1} = c_1, x_{i2} = c_2, \dots, x_{ik} = c_k) = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k$$

e o estimador de θ é

$$\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k.$$

Fazendo $c = [1 \quad c_1 \quad \dots \quad c_k]$, logo o valor esperado e variância do estimador de θ são os seguintes:

$$E(\hat{\theta}|X, c) = E(\hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k|X, c) = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k = \theta$$

$$Var(\hat{\theta}|X, c) = Var(c\hat{\beta}|X, c) = c\, Cov(\hat{\beta}|X, c)c^T = \sigma^2 c(X^T X)^{-1} c^T$$

Portanto, **erro padrão da previão em média** é dado por

$$s_{\hat{\theta}} = \hat{\sigma}\sqrt{c(X^T X)^{-1} c^T}$$

e o respectivo I.C. de previsão para θ com confiança $(1 - \alpha)$ é dado por

$$\left[\hat{\theta} - t_{1-\alpha/2} s_{\hat{\theta}}\,;\, \hat{\theta} + t_{1-\alpha/2, n-2} s_{\hat{\theta}}\right].$$

50

b) Previsão pontual (previsão individual)

Considere-se de novo que $x_{i1} = c_1, x_{i2} = c_2, \dots, x_{ik} = c_k$, e seja

$$y_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k + \varepsilon_0$$

Considere-se o previsor mínimo quadrado de y_0,

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$$

e o erro de previsão,

$$\varepsilon = y_0 - \hat{y}_0.$$

Utilizando a variável aleatória ε, vão estudar-se as propriedades estatísticas do previsor. Como

$$E(\varepsilon|X, c) = E(y_0 - \hat{y}_0|X, c) = 0$$

a variância de ε, condicionada por X e c, é dada por

$$Var(\varepsilon|x, c) = \sigma^2 \left\{1 + c(X^T X)^{-1} c^T\right\}$$

Assim, o **erro padrão da previsão** é dado por

$$s_\varepsilon = \hat{\sigma}\sqrt{1 + c(X^T X)^{-1} c^T}$$

E o respetivo I.C. de previsão para y_0 com confiança $(1 - \alpha)$ é dado por

$$\left[\hat{y}_0 - t_{1-\alpha/2} s_\varepsilon\,;\, \hat{y}_0 + t_{1-\alpha/2, n-2} s_\varepsilon\right].$$

51

Exemplo 1

Foram controladas as variáveis x_1 e x_2, no intuito de verificar se elas explicam as variações de y, tendo sido obtidos os resultados seguintes:

x_1	1	4	9	11	3	8	6	13	2	7	6
x_2	8	2	-8	-10	6	-6	0	-13	4	-2	-4
y	6	8	1	0	5	3	2	-4	10	-3	5

a) Ajuste o modelo da forma $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$.

b) Calcule as estimativas de $var(\hat{\boldsymbol{\beta}})$ e determine o I.C. dos parâmetros do modelo.

c) Determine a soma dos quadrados total (SQ_T) construa a tabela ANOVA. Será que se pode considerar a regressão significativa?

d) Qual a percentagem de variação de y explicada linearmente por x_1 e x_2?

e) Considere-se $[\boldsymbol{X}^T\boldsymbol{X}]^{-1} = \begin{bmatrix} 4.3708 & -0.8156 & -0.4355 \\ -0.8156 & 0.1558 & 0.0840 \\ -0.4355 & 0.0840 & 0.0475 \end{bmatrix}$;

Calcule uma previsão (média e pontual) para $x_{i1} = 5$ e $x_{i2} = 3$ determine os respetivos I.C.

52

Para resolver as questões anteriores, considere os seguintes output do *software* de estatística R.

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0117 -1.0575  0.3153  1.0784  2.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.4878     4.2692   4.565  0.00184 **
X1           -2.9628     0.8060  -3.676  0.00626 **
X2           -1.1318     0.4451  -2.543  0.03455 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.042 on 8 degrees of freedom
Multiple R-squared:  0.8244,   Adjusted R-squared:  0.7805
F-statistic: 18.78 on 2 and 8 DF,  p-value: 0.0009504
```

a) O modelo ajustado é dado por:

$y_i = 19.4878 - 2.9628x_{i1} - 1.1318x_{i2} + \varepsilon_i$

b) As estimativas de $var(\hat{\boldsymbol{\beta}})$ são

- $var(\hat{\beta}_0) = 18{,}23$
- $var(\hat{\beta}_1) = 0{,}65$
- $var(\hat{\beta}_2) = 0{,}20$

I.C. dos parâmetros $\hat{\boldsymbol{\beta}}$:

- I.C. de $\hat{\beta}_0$: [9,64 ; 29,33]
- I.C. de $\hat{\beta}_1$: [-4,82 ; -1,10]
- I.C. de $\hat{\beta}_2$: [-2,16 ; -0,11]

53

c) Tabela Anova

F.V.	SQ	gl	MQ	F_{obs}
Reg.	156,64	2	78,32	18,78
Res.	33,36	8	4,17	
Tot.	190	10		

```
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X           2 156.64   78.32  18.782 0.0009504 ***
Residuals  8  33.36    4.17
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A regressão significativa a 5%, uma vez que o $p - value = 0{,}00095 < \alpha = 0{,}05$.

d) 82,44% da variabilidade de y é explicada pelo modelo linear baseado em x_1 e x_2

e) Previsão para $x_{i1} = 5$ e $x_{i2} = 3 \implies \hat{y}_0 = 19.4878 - 2.9628 \times 5 - 1.1318 \times 3 = 1{,}2784$

- I.C. para previsão em média: [-1,8627 ; 4,4195]
- I.C. para previsão pontual: [-4,3821 ; 6,9389]

54

4.5. Validação dos pressupostos do modelo - análise de resíduos

Os pressupostos do M.R.L.M. são basicamente os mesmos apresentados para o M.R.L.S., e ainda: colinearidade (isto é, os regressores devem ser independentes)

4.5.1. Diagnóstico de pontos influentes

As observações influentes são aquelas que individualmente ou em conjunto com as outras observações demonstram ter mais impacto do que as restantes no cálculo das estimativas dos parâmetros do modelo.

55

As observações influentes, são habitualmente identificadas como as que violam, pelo menos um, dos seguintes critérios.

A observação $x_i = (x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$ é influente se

- $DFBETA_{j(i)}$: $\left|\frac{\hat{\beta}_i - \hat{\beta}_{j(i)}}{\sqrt{QME_{(i)} c_{jj}}}\right| > \frac{2}{\sqrt{n}}$
- $DFFITS_i$: $\left|\frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{QME_{(i)} h_{ii}}}\right| > 2\sqrt{\frac{k+1}{n-p}} \longrightarrow p$ é número de parâmetro
- Distância Cook: $D_i = \frac{e_i^2 h_{ii}}{(k+1) QME (1-h_{ii})^2} > 1$

onde,

- $\hat{y}_{i(i)}$ é previsão de y_i removendo a observação i,
- $QME_{(i)} = \sigma^2_{(i)}$ é variância do erro quando removendo a observação i,
- h_{ii} é diagonal principal da matriz chapéu, $H = \boldsymbol{X}[\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T$,
- $\hat{\beta}_{j(i)}$ é o valor estimado para β_j quando excluímos a observação (x_i, y_i).

56

Exemplo 2 _ Retornam-se ao exemplo 1

Considere o output do *software* R abaixo, verifique se existem pontos influentes no modelo

```
Influence measures of
     lm(formula = Y ~ X) :

    dfb.1_  dfb.X1  dfb.X2   dffit cov.r  cook.d   hat inf
1  -0.0352 -0.0325 -0.1352 -0.5761 1.614 0.11450 0.313
2   0.1801 -0.0982 -0.0313  0.5744 0.760 0.09575 0.131
3  -0.1275  0.1129  0.1408 -0.2400 1.737 0.02130 0.214
4  -0.1373  0.1699  0.0743  0.5794 1.321 0.11165 0.248
5  -0.3052  0.3659  0.4321  0.5777 1.967 0.11829 0.389
6   0.0508 -0.0432 -0.0512  0.0923 1.741 0.00322 0.159
7  -0.0410  0.0495  0.0518  0.0718 1.829 0.00196 0.191
8  -0.0703  0.0763  0.0466  0.1701 2.610 0.01096 0.437   *
9   0.3331 -0.2969 -0.2477  0.4143 2.020 0.06257 0.352
10  0.7563 -0.9146 -0.8965 -1.3696 0.135 0.30216 0.164
11 -0.5748  0.5423  0.5485 -0.6246 1.962 0.13721 0.401
```

Critérios:

- $\left|DFBETA_{j(i)}\right| > \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{11}} = 0.6030$
- $|DFFITS_i| > 2\sqrt{\frac{k+1}{n-p}} = 2 \times \sqrt{\frac{3}{11-3}} = 1.2247$
- $D_i > 1$

Assim, a observação 10 é um ponto influente no modelo de regressão linear construído.

57

4.5.2. Colinearidade

Existe colinearidade quando, no modelo regressão linear, as variáveis independentes estão fortemente correlacionadas entre si.

A colinearidade pode ser diagnosticada pela análise da matriz correlação bivariada

$$r = \begin{bmatrix} r_{(x_1,x_1)} & \cdots & r_{(x_1,x_k)} \\ \vdots & \ddots & \vdots \\ r_{(x_k,x_1)} & \cdots & r_{(x_k,x_k)} \end{bmatrix}; \text{ em que } r_{x_j,x_k} = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{(n-1)s_{x_j}s_{x_k}}; i = 1, \ldots, n;$$

s_{x_j} e s_{x_k} são desvio padrão de X_j e X_k respetivamente

- Se existirem apenas duas variáveis independentes, é fortemente correlacionados, então a seleção de um bom modelo pode ser feita do seguinte modo:
  - Testar a significância do modelo linear que inclui cada uma das variáveis independentes separadamente (ou testar dois M.R.L.S.).
  - Observar o valor de coeficiente de determinação e coeficiente de determinação ajustado do modelo e escolher o modelo com melhor desempenho.

58

Exemplo 3 _ Retornam-se ao exemplo 1

Considere output do *software* R. Verifique se existe multicolinearidade no modelo e indique um modelo adequado para ajustar aos dados.

Matriz correlação

```
> cor(da)
             Y          X1          X2
Y    1.0000000 -0.8261303  0.7265308
X1  -0.8261303  1.0000000 -0.9768966
X2   0.7265308 -0.9768966  1.0000000
```

```
Call:
lm(formula = Y ~ X1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3887 -1.2901  0.5323  1.6113  2.8085

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.1125     1.5940   5.717 0.000288 ***
X1           -0.9605     0.2184  -4.398 0.001724 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.589 on 9 degrees of freedom
Multiple R-squared:  0.6825,  Adjusted R-squared:  0.6472
F-statistic: 19.35 on 1 and 9 DF,  p-value: 0.001724
```

```
Call:
lm(formula = Y ~ X2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0424 -1.8427  0.6893  2.3570  4.1588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9753     1.0003   3.974  0.00324 **
X2            0.4665     0.1471   3.172  0.01133 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.157 on 9 degrees of freedom
Multiple R-squared:  0.5278,  Adjusted R-squared:  0.4754
F-statistic: 10.06 on 1 and 9 DF,  p-value: 0.01133
```

Resolução

- Existe colineariedade, pois x_1 e x_2 têm uma correlação negativa forte ($r \approx -0{,}9769$).
- O bom modelo é o modelo completo, uma vez que as variáveis x_1 e x_2 são estatisticamente significativas (significância 5%) e no modelo $y_i = 9{,}1125 - 0{,}9605x_{i1} + \varepsilon_i$ ($y_i = 3{,}9753 + 0{,}4665x_{i2} + \varepsilon_i$) a variabilidade explicada pelo modelo é cerca de 68,25% (52,78%) com grau de ajustamento de 64,75% (47,54%).
  Logo o modelo considerar é $y_i = 19.4878 - 2.9628x_{i1} - 1.1318x_{i2} + \varepsilon_i$.

59

- Para mais do que dois regressores, a colinearidade pode ser diagnosticada através do fator de inflação da variância (VIF),

$$VIF_j = \frac{1}{1-R_j^2};$$

onde R_j^2 é o R^2 da regressão de X_j sobre as outras variáveis explicativas

Geralmente, a colinearidade existe quando $VIF \geq 5$. (Mello, F. M. 2014)

Numa situação destas deve-se eliminar uma das variáveis e reestimar os parâmetros do modelo.
Uma forma sistemática de executar este procedimento é o que se apresenta de seguida.

60

4.6. Seleção de variáveis numa regressão múltipla

Um método amplamente utilizado é o ***Stepwise,*** que consiste na combinação dos métodos *Backward (inclusão passo atrás)* e *Forward* (inclusão passo a frente). Geralmente parte-se de um modelo completo e, em cada passo avalia-se a exclusão e a inclusão de variáveis.

Um critério muito utilizado para a comparação de modelos é o *Akaike information Criterion* (AIC) dado por

$$AIC = -2\ln(L|\beta) + 2p$$

em que $\ln(L|\beta)$ é o log natural da função de verossimilhança do modelo e p é o número de parâmetros do modelo. Quanto menor o valor do AIC, melhor é o ajuste do modelo aos dados.

NOTA: Neste disciplina, a seleção de variáveis é feita através do *Software R*.

61

Exemplo 4

Uma experiência foi conduzida para estudar o tamanho de lulas utilizadas como "isco" para tubarões. As variáveis regressoras $x_1, \ldots, x_5$ representam características das lulas (medidas e pesos), enquanto a resposta considerada traduz o peso do tubarões (kg) em 22 tubarões.

y	0,90	1,33	0,33	0,37	0,50	0,56	0,47	0,89	0,29	0,96	0,91	0,87	3,93	2,06	3,90	2,83	3,46	2,92	2,50	3,57	4,66	3,16
x_1	3,34	3,95	2,53	2,53	2,68	2,78	2,75	3,24	2,53	3,42	3,32	3,39	4,74	4,03	1,50	4,59	4,46	4,39	4,28	4,46	5,59	4,41
x_2	2,73	3,80	2,14	2,12	2,30	2,37	2,12	2,75	2,17	2,88	2,81	2,81	3,75	3,42	3,06	3,98	4,03	3,65	4,00	4,06	2,74	4,26
x_3	1,12	1,35	0,87	0,87	0,92	1,07	1,02	1,12	0,12	1,15	1,15	1,22	1,53	1,33	1,71	1,68	1,61	1,63	1,84	1,73	1,91	1,61
x_4	1,91	2,30	1,45	1,38	1,63	1,56	1,30	1,96	1,43	1,96	1,94	1,96	2,58	2,42	4,06	2,60	2,78	2,60	2,45	2,75	3,16	2,91
x_5	0,89	1,20	0,82	0,69	0,77	0,79	0,79	0,87	0,74	0,94	0,97	0,97	1,66	1,28	5,02	1,50	1,50	1,61	1,73	1,58	1,84	1,40

62

Para o conjunto de dados referido anteriormente obteve-se o seguinte output do R.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5931 -0.2131 -0.0848  0.2842  0.5822

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.0149     0.5214  -5.782  2.8e-05 ***
X1            0.3978     0.5902   0.674   0.5099
X2           -0.2788     0.2282  -1.222   0.2394
X3           -0.3534     0.4577  -0.772   0.4512
X4            1.1958     0.7878   1.518   0.1486
X5            1.7940     0.7778   2.307   0.0348 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3661 on 16 degrees of freedom
Multiple R-squared:  0.9505,   Adjusted R-squared:  0.935
F-statistic: 61.41 on 5 and 16 DF,  p-value: 7.156e-10
```

a) Estabeleça o modelo de regressão a ajustar aos dados.

b) Verifique se todos os coeficientes são estatisticamente significativos ao nível de significância 5%.

63

```
Start:  AIC=-39.22
Y ~ X1 + X2 + X3 + X4 + X5

       Df Sum of Sq    RSS     AIC
- X1    1   0.06091 2.2054 -40.602
- X3    1   0.07993 2.2245 -40.414
- X2    1   0.20013 2.3447 -39.256
<none>              2.1445 -39.219
- X4    1   0.30879 2.4533 -38.259
- X5    1   0.71309 2.8576 -34.903

Step:  AIC=-40.6
Y ~ X2 + X3 + X4 + X5

       Df Sum of Sq    RSS     AIC
- X3    1   0.05728 2.2627 -42.038
<none>              2.2054 -40.602
+ X1    1   0.06091 2.1445 -39.219
- X2    1   0.40079 2.6062 -38.929
- X5    1   1.63395 3.8394 -30.406
- X4    1   2.12766 4.3331 -27.745

Step:  AIC=-42.04
Y ~ X2 + X4 + X5

       Df Sum of Sq    RSS     AIC
<none>              2.2627 -42.038
+ X3    1   0.05728 2.2054 -40.602
+ X1    1   0.03826 2.2245 -40.414
- X2    1   0.44514 2.7079 -40.087
- X5    1   1.69294 3.9557 -31.750
- X4    1   2.08311 4.3458 -29.680
```

```
Call:
lm(formula = Y ~ X2 + X4 + X5)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54769 -0.24121 -0.02671  0.30193  0.72159

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.7426     0.3392  -8.085 2.11e-07 ***
X2           -0.3660     0.1945  -1.882 0.076138 .
X4            1.5831     0.3889   4.071 0.000717 ***
X5            1.9512     0.5317   3.670 0.001753 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3546 on 18 degrees of freedom
Multiple R-squared:  0.9477,   Adjusted R-squared:  0.939
F-statistic: 108.8 on 3 and 18 DF,  p-value: 9.999e-12
```

c) Considere a seleção das variáveis usando o método Stepwise indique as variáveis independentes que foram selecionadas para incluir no modelo. Escreva o novo modelo de regressão a ajustar aos dados.

64

Resolução

a) O modelo ajustado inicialmente

$$y_i = -3{,}0149 + 0{,}3978x_{i1} - 0{,}2788x_{i2} - 0{,}353x_{i3} + 1{,}1958x_{i4} + 1{,}7940x_{i5} + \varepsilon_i$$

b) A constante e o coeficiente de x_1 são estatisticamente significativos $(p - value \leq 5\%)$ enquanto que os restantes coeficientes não são significativos $(p - value > 5\%)$.

c) O resultado do método Stepwise mostra que as variáveis independentes escolhidas para incluir no modelo são x_2, x_4 e x_5 e portanto o modelo é

$$y_i = -2{,}7426 - 0{,}3660x_{i2} + 1{,}5831x_{i4} + 1{,}9512x_{i5} + \varepsilon_i$$

65 **Exercícios**

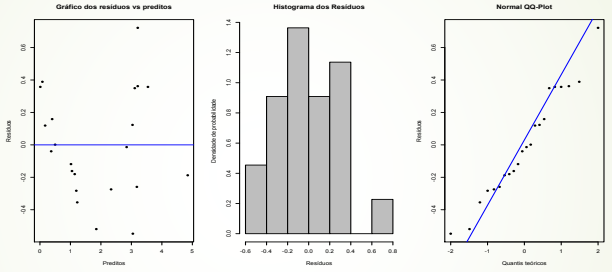
1. Proceda à análise dos resíduos por forma a validar os pressupostos do modelo para o exemplo 3.
2. Baseado no resultado do exemplo 4 alínea c:
   i. Calcule as estimativas de $var(\hat{\boldsymbol{\beta}})$ e determine o I.C. dos parâmetros do modelo.
   ii. Completa os espaços a tracejado, na tabela ANOVA a seguir:

F.V.	$\boldsymbol{SQ}$	$\boldsymbol{gl}$	$\boldsymbol{MQ}$	$\boldsymbol{F_{obs}}$	$p-value$
Reg.	41,0363	...	...	....	0,00095
Res.	2,2627	...	...		
Tot.	...	...	...		

   iii. Considere os resultados da tabela ANOVA na alínea anterior:
      a) Determine a percentagem de variabilidade das observações que é explicada pelo modelo linear e quantificar o grau de ajustamento do modelo linear.
      b) Em nível de significância 5%, será que o modelo é estatisticamente significativo? Porque ?

66

   iv. Através de visualização dos gráficos e o output do R a seguir, comenta o que é que pode concluir sobre a validação dos pressupostos do modelo.




```
    Shapiro-Wilk normality test

data:  rstudent(modelo2)
W = 0.97958, p-value = 0.9093
```

```
    Durbin-Watson test

data:  modelo2
DW = 2.5361, p-value = 0.8709
alternative hypothesis: true autocorrelation is greater than 0
```

   v. Construa um I.C. de 95% para uma previsão para y através de $x_2 = 3$, $x_4 = 1$ e $x_5 = 2{,}5$.

67

Soluções:

1. Os pressupostos do modelo são válidas.
2. i. As estimativas de $var(\widehat{\boldsymbol{\beta}})$:
   - $var(\hat{\beta}_0) = 0{,}1151$
   - $var(\hat{\beta}_2) = 0{,}0378$
   - $var(\hat{\beta}_4) = 0{,}1512$
   - $var(\hat{\beta}_5) = 0{,}2827$

   I.C. dos parâmetros $\widehat{\boldsymbol{\beta}}$:
   - I.C. de $\hat{\beta}_0$: [-3,4552 ; -2,0300]
   - I.C. de $\hat{\beta}_2$: [-0,7746 ; 0,0426]
   - I.C. de $\hat{\beta}_4$: [0,7661 ; 2,4002]
   - I.C. de $\hat{\beta}_5$: [0,8341 ; 3,0682]

   iii. a). 94,77% variabilidade de observação é explicada pelo modelo linear, e o grau de ajustamento de modelo é 93,9%.

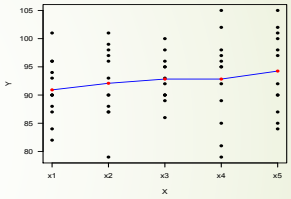
   v. Previsão
   - I.C. para previsão em média: [0,3292 ; 4,9119]
   - I.C. para previsão pontual: [0,2112 ; 5,0299]

68

5. Análise de variância (ANOVA)

ANOVA é uma metodologia estatística com o objetivo de comparar de três ou mais grupos, quer no que respeita à sua localização. Neste caso, é utilizada para verificar se existem diferenças significativas entre as médias dos grupos, que sejam resultado dos efeitos dos grupos.

Gráfico de médias



- $X \rightarrow$ fator
- $x_j; j = 1,..,k$ (tratamentos/grupos)
- $x_j \rightarrow$ fixos ou aleatório
- $n \times k$ observações

69

Fator X:

- **Efeitos fixos:** os grupos são determinados à partida.
- **Efeitos aleatórios**: os grupos foram selecionadas aleatoriamente dum grande número (infinito) de população.

Pressupostos básicos da ANOVA a um fator (One-way ANOVA)

- As amostras são aleatórias e independentes.
- As populações têm distribuição normal (o teste é paramétrico).
- As variâncias populacionais são iguais (homogeneidade de variância entre grupos).

Tipos de ANOVA

- ANOVA paramétrica
  - *ANOVA* com um fator e efeitos fixos (*ANOVA one-way and fixed effects*)
  - *ANOVA* com um fator e efeitos aleatórios (*ANOVA one-way and random effects*)
- ANOVA não paramétrica

70

5.1. *ANOVA* com um fator e efeitos fixos (*ANOVA one-way and fixed effects*)

Neste caso, cada observação é expressa como a soma da média geral μ e o efeito principal α_j, que é fixo no sentido de que a população com média $\mu_j = \mu + \alpha_j$ é fixada pelo investigador.

- **Modelo**: $Y_{ij} = \mu_j + \varepsilon_{ij} = \mu + \alpha_j + \varepsilon_{ij}; i = 1, \dots, n_j; j = 1, \dots, k$

  onde, $Y_{ij} \longrightarrow$ observação do i do grupo j;

  $\mu_j \longrightarrow$ média do grupo j;

  $\alpha_j \longrightarrow$ efeito (não aleatório) do grupo j;

  $\varepsilon_{ij} \longrightarrow$ erro aleatório da observação do i do grupo j.

- Hipóteses a testar

  $H_0: \mu_1 = \mu_2 = \cdots = \mu_k = \mu$

  $H_1: \mu_j \neq \mu$, para algum j

71

- Estatística do teste: $F = \frac{MQ_f}{MQ_E} \sim F_{k-1,n-k}$

Tabela ANOVA

Fonte de variação	SQ	$g.l.$	MQ	F_{obs}	$valor - P$
Fatores	SQ_f	$k-1$	$MQ_f = \frac{SQ_f}{k-1}$	$\frac{MQ_f}{MQ_E}$	$P(F > F_{obs})$
Erros	SQ_E	$n-k$	$MQ_E = \frac{SQ_E}{n-k}$		
Total	SQ_T	$n-1$			

$$SQ_f = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 \; ; \; SQ_E = \sum_{j=1}^{k} (x_{ij} - \bar{x}_j)^2 \; ; \; SQ_T = SQ_f + SQ_E$$

"Tamanho do efeito" $= \eta^2 = \frac{SQ_f}{SQ_T}$

Onde, $\bar{x}$ ⟶ média global
$\bar{x}_j$ ⟶ média do grupo j
k ⟶ # grupos (níveis do fator)
n_j ⟶ # observações do grupo j
n ⟶ # total de observações

Decisão: Rejeitar H_0 (em favor de H_1) se $F_{obs} > F_{1-\alpha,k-1,\,n-k}$. Caso contrário não se rejeitar H_0.

72

No caso de rejeitar H_0 é desejável efetuar comparação de médias de grupos duas a duas por forma a detetar diferenças estatisticamente significativas **(comparações múltiplas).**
O número total de comparações é $m = C_2^k = \frac{k!}{2!(k-2)!}$.

Hipóteses a testar em cada comparação:

$H_0: \mu_i = \mu_j$
$H_1: \mu_i \neq \mu_j$; $\quad i \neq j; \quad i, j = 1, \ldots, k$

Diferença mínima significativo
(ou H.S.D. – Honestly significant difference)

Região de rejeição

- **Teste de Tukey :** $|\bar{X}_i - \bar{X}_j| \geq \frac{Q_{\alpha;k;n-k}\sqrt{MQ_E}}{\sqrt{n_j}}$ onde $Q_{\alpha;k;n-k}$ é o quantil de probabilidade $(1-\alpha)$ para distribuição Studentized Range (tabelada) com $(k,\ n-k)$ graus de liberdade.
  (amostras de igual dimensão)

- **Teste de Scheffé** :$|\bar{X}_i - \bar{X}_j| > \sqrt{MQ_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)(k-1)F_{1-\alpha,k-1,\,n-k}}$ onde $F_{1-\alpha,k-1,\,n-k}$é o quantil de probabilidade $(1-\alpha)$ para distribuição F de Snedecor (tabelada) com $(k-1,\ n-k)$ graus de liberdade.
  (para qualquer dimensão)

73 **Exemplo 1**

Um grupo de alunos de uma escola foram sujeitos a quatro técnicas diferentes de ensino. Ao fim de certo tempo foram testados obtendo-se os resultados da tabela:

1ª	2ª	3ª	4ª
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88
81	72	83	
69	79	76	
	90		

a) Verifique se existem diferenças significativas entre as médias fornecidas pelas técnicas de ensino.
b) Se existirem diferenças significativas, identifique os grupos que se destacam dos restantes.

74 **Resolução:**

i	1ª	2ª	3ª	4ª
1	65	75	59	94
2	87	69	78	89
3	73	83	67	80
4	79	81	62	88
5	81	72	83	
6	69	79	76	
7		90		
n_j	6	7	6	4
n	23			
$\sum_{i=1}^{n_j} x_{ij}$	454	549	425	351
$\bar{x}_j$	75,67	78,43	70,83	87,75
$\bar{x}$	77,35			

$$SQ_f = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2 = 6\times(75{,}67 - 77{,}35)^2 + \cdots + 4\times(87{,}75 - 77{,}35)^2 = 712{,}59$$

$$SQ_E = \sum_{j=1}^{k} (x_{ij} - \bar{x}_j)^2 = (65 - 75{,}67)^2 + \cdots + (69 - 75{,}67)^2 + (75 - 78{,}43)^2 + \cdots$$
$$+(90 - 78{,}43)^2 + \cdots + (88 - 87{,}75)^2 = 1196{,}63$$

$$SQ_T = SQ_f + SQ_E = 1909{,}22$$

75

a) Hipóteses a testar:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$

$H_1: \mu_j \neq \mu$, para algum $j = 1, \ldots, 4$.

Estatística do teste: $F = \frac{MQ_f}{MQ_E}$

Tabela ANOVA

Fonte de variação	SQ	$g.l.$	MQ	F_{obs}
Fatores	$SQ_f = 712{,}59$	3	$MQ_f = 237{,}53$	3,77
Erros	$SQ_E = 1196{,}63$	19	$MQ_E = 62{,}98$	
Total	$SQ_T = 1909{,}22$	22		

Ponto crítico: $F_{1-\alpha, k-1, n-k} = F_{0{,}95, 3, 19} = 3{,}13$

Decisão: Como $F_{obs} > F_{0{,}95, 3, 19}$, (3,77 > 3,13), então rejeita-se H_0 em favor de H_1. Assim, existem diferenças significativas nas médias fornecidas pelas técnicas de ensino ($\alpha = 5\%$).

76

b) Teste de Scheffé

Hipóteses a testar:

$H_0: \mu_i = \mu_j$

$H_1: \mu_i \neq \mu_j, i \neq j; i,j = 1, \ldots, 4$; para algum j

- Para X_1 e X_2

  Valor crítico: $\sqrt{MQ_E \left(\frac{1}{n_1} + \frac{1}{n_2}\right)(k-1)F_{0{,}95, 3, 19}} = \sqrt{62{,}98 \times \left(\frac{1}{6} + \frac{1}{7}\right) \times 3 \times 3{,}13} = 13{,}53$

  $|\bar{X}_1 - \bar{X}_2| = |75{,}67 - 78{,}43| = 2{,}76 < 13{,}53 \Longrightarrow$ não se rejeita H_0 em favor de H_1.

  (. . .) Assim sucessivamente para todos os pares das médias

- Para X_3 e X_4

  Valor crítico: $\sqrt{MQ_E \left(\frac{1}{n_3} + \frac{1}{n_4}\right)(k-1)F_{0{,}95, 3, 19}} = \sqrt{62{,}98 \times \left(\frac{1}{6} + \frac{1}{4}\right) \times 3 \times 3{,}13} = 15{,}697$

  $|\bar{X}_3 - \bar{X}_4| = |70{,}83 - 87{,}75| = 16{,}92 > 13{,}53 \Longrightarrow$ rejeitar-se H_0 em favor de H_1.

Assim, pode concluir-se que existem diferenças significativas entre as médias μ_3e μ_4. ($\alpha = 5\%$).

77

5.2. *ANOVA* com um fator e efeitos aleatórios (*ANOVA one-way and random effects*)

- **Modelo**: $Y_{ij} = \mu_j + \varepsilon_{ij} = \mu + \alpha_j + \varepsilon_{ij}; i = 1, \dots, n_j; j = 1, \dots, k.$

  Neste caso, α_j são variáveis aleatórias onde $\alpha_j \sim N(0, \sigma_A^2)$ e σ_A^2 é a variância do fator, adicionalmente $\varepsilon_{ij} \sim N(0, \sigma^2)$ onde σ^2 é a variância do erro.

- Hipóteses a testar

  $H_0: \sigma_A^2 = 0$ (todos os grupos são idênticos)

  $H_1: \sigma_A^2 > 0$ (existe variação entre grupos)

- Estatística do teste: $F = \frac{MQ_f}{MQ_E}$

Decisão: Rejeitar H_0 (em favor de H_1) se $F_{obs} > F_{1-\alpha, k-1, n-k}$. Caso contrário não se rejeitar H_0.

Quando H_0 é rejeitada, faz sentido estimar a variância do fator:

- Para grupos com o mesmo dimensão ($n_1 = n_2 = \cdots = n_k$), $\sigma_A^2 = \frac{MQ_f - MQ_E}{n_j}$
- Caso os grupos tiverem dimensões diferentes, n_j deve ser substituir por $r = \frac{1}{k-1}\left[\sum_{j=1}^{k} n_j - \frac{\sum_{j=1}^{k} n_j^2}{\sum_{j=1}^{k} n_j}\right]$.

78

Exemplo 2

Efetuou-se uma experiência com o objetivo de investigar a variabilidade do latex obtido de árvore de borracha numa grande plantação. Para o efeito, escolheram-se aleatoriamente cinco árvore da plantação. De cada árvore retiram-se, aleatoriamente, 7 amostras de latex e mediu-se a sua elasticidade, tendo-se obtido os resultados seguintes:

	Árvore			
I	**II**	**III**	**IV**	**V**
5,07	5,20	4,52	5,15	5,08
5,20	4,50	4,96	5,39	4,74
4,51	5,29	4,80	4,90	4,92
5,33	5,36	3,72	5,20	4,17
4,97	4,80	4,50	4,99	4,79
5,04	5,58	4,80	4,78	5,77
5,29	5,06	4,03	5,37	4,49

a) Será que a elasticidade varia de árvore para árvore, na plantação ?

b) Estime a variância do fator, se existir a variabilidade.

79

Resolução

a) Hipóteses a testar

$H_0: \sigma_A^2 = 0$ vs. $H_1: \sigma_A^2 > 0$

Estatística de teste: $F = \frac{MQ_f}{MQ_E}$

Tabela ANOVA

F.V.	SQ	$g.l.$	MQ	F_{obs}
Fator	$SQ_f = 2{,}07$	4	$MQ_f = 0{,}52$	3,60
Erros	$SQ_E = 4{,}30$	30	$MQ_E = 0{,}14$	
Total	$SQ_T = 6{,}37$	34		

Ponto crítico: $F_{1-\alpha,k-1,\,n-k} = F_{0{,}95,4,30} = 2{,}69$

Decisão: Como $F_{obs} > F_{0{,}95,4,30}$, (3,60 > 2,69) então rejeita-se H_0 em favor de H_1. Assim, conclui-se que a elasticidade média varia de árvore para árvore, na plantação.

b) A variância do fator é dada por $\sigma_A^2 = \frac{MQ_f - MQ_E}{n} = \frac{0{,}52-0{,}14}{7} \approx 0{,}05$

80

5.3. Validação dos pressupostos da ANOVA a um fator

- Normalidade dos dados em cada grupo $j = 1,1,\ldots,k$ (como já referido no modelo de regressão linear)
  - Teste de Kolmogorov Smirnov (KS)
  - Teste de Shapiro Wilk
  - QQ-plot
- Homogeneidade de variâncias
  - Teste de Bartlett (assume normalidade dos dados)
  - Teste de Levene (não assume normalidade dos dados)

81

- **Teste de Bartlett**

É um teste paramétrico para comparação de duas ou mais variâncias populacionais.

Hipóteses a testar

H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

H_1: $\sigma_i^2 \neq \sigma_j^2; i \neq j; i, j = 1, \ldots, k$; para algum j

Estatística do teste

$$B = \frac{2{,}3026\left[\sum_{j=1}^{k}(n_j-1).ln\left(\frac{\sum_{j=1}^{k}(n_j-1)s_j^2}{\sum_{j=1}^{k}(n_j-1)}\right) - \sum_{j=1}^{k}(n_j-1).lns_j^2\right]}{1+\frac{1}{3(k-1)}.\left[\sum_{j=1}^{k}\frac{1}{n_j-1} - \frac{1}{\sum_{j=1}^{k}(n_j-1)}\right]} = \frac{C}{A} \sim \chi^2_{k-1}$$

onde, n_j é dimensão do grupo j e s_j^2 é variância do grupo j.

Decisão: rejeita-se H_0 quando $B > \chi^2_{(1-\alpha);(k-1)}$. Quando $C < \chi^2_{(1-\alpha);(k-1)}$ não é necessário calcular A, podendo logo concluir-se que não se rejeita H_0 em favor de H_1.

82

- **Teste de Levene**

É um teste paramétrico para testar a homogeneidade das variâncias.

Hipóteses a testar

H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

H_1: $\sigma_i^2 \neq \sigma_j^2; i \neq j; i, j = 1, \ldots, k$; para algum j

Estatística do teste

$$W = \frac{n-k}{k-1} \times \frac{\sum_{j=1}^{k} n_j\left(\bar{Z}_j - \bar{Z}\right)^2}{\sum_{j=1}^{k}\sum_{i=}^{n_j}\left(Z_{ij} - \bar{Z}_j\right)^2} \sim F_{k-1,n-k}$$

Onde, $Z_{ij} = |x_{ij} - \bar{X}_j|; i = 1, \ldots, n_j; j = 1, \ldots, k$;

$x_{ij} \longrightarrow$ observação i do grupo j;

$\bar{X}_j \longrightarrow$ média do grupo j;

$\bar{Z}_j \longrightarrow$ média dos valores de Z para o grupo j;

$\bar{Z} \longrightarrow$ média global dos valores de Z.

Se a variável não têm distribuição normal, então Z deve calcular-se por $Z_{ij} = |x_{ij} - \tilde{X}_j|$ onde $\tilde{X}_j$ é a mediana do grupo j.

Decisão: rejeita-se H_0 quando $W \geq F_{1-\alpha;k-1,n-k}$. Caso contrário não se rejeita H_0.

83

Exemplo 3

Baseada no enunciado da questão do exemplo 1 e exemplo 2, avaliar se os pressupostos da análise são validos.

Resoluções

Do exemplo 1 e 2 – Normalidade (**Exercício**)

- Do exemplo 1 – Homogeneidade

  Hipóteses a testar

  H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$

  H_1: $\sigma_i^2 \neq \sigma_j^2; i \neq j; i, j = 1, \dots, 4$; para algum j

  O teste foi feita através do R:

```
    Bartlett test of homogeneity of variances

data:  Nota by Tecnica
Bartlett's K-squared = 0.94217, df = 3, p-value = 0.8152
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  1.2125 0.3323
      19
```

  Dos testes feitos através do R, conclui-se não rejeitam H_0 em favor de H_1 $(\alpha = 5\%)$. Portanto, considera-se o pressupostos ANOVA de homogeneidade de variâncias como válido.

- Do exemplo 2 – Homogeneidade (**Exercício**)

84

5.4. Análise de variância não paramétrica – teste de Kruskal-Wallis

É a alternativa da one-way ANOVA, quando os pressupostos de normalidade e homogeneidade são violados.

Este teste permite encontrar diferenças significativas entre os valores centrais (mediana) de 3 ou mais amostras independentes. Aplica-se a variáveis de nível pelo menos ordinal, sendo baseado em ordenações.

Antes de calcular a estatística de teste ordenar todas as observações por ordem crescente e a atribuindo a cada uma a sua ordem na amostra global, e no caso empates (haver os valores repetidos) atribui-se a média das ordens que as observações teriam.

Hipóteses a testar

H_0: $\theta_1 = \theta_2 = \cdots = \theta_k$

H_1: $\theta_i \neq \theta_j; i \neq j; i, j = 1, \dots, k$, para algum j

85 Estatística de teste

$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j{}^2}{n_j} - 3(n+1)$ (sem valores repetidos nas amostras)

$H^* = \frac{H}{1 - \frac{\sum_{j=1}^{k}(t_i^3 - t_i)}{n^3 - n}}$ (com valores repetidos nas amostras)

onde, $R_j = \sum_{i=1}^{n_j} r_{ij}$ ⟶ soma das ordens do grupo j;

r_{ij} ⟶ ordem da observação i da grupo j;

t_i ⟶ # observações repetidos.

Os valores críticos para a rejeição da H_0 encontram-se tabelados, mas nos casos

- $k = 3$ e $n_j \geq 6; j = 1, 2, 3$
- $k > 3$ e $n_j \geq 5; j = 1, 2, \dots, k$

A distribuição de H é tal que $H \sim \chi^2_{k-1}$

Decisão: rejeita-se H_0 em favor de H_1 se $H_{obs} > \chi^2_{1-\alpha;k-1}$. Caso contrário não se rejeita H_0.

86 No caso de rejeitar H_0 é desejável efetuar um procedimento de comparação múltipla por forma a detetar diferenças estatisticamente significativas entre si.

Hipótese a testar

$H_0: \theta_i = \theta_j$

$H_1: \theta_i \neq \theta_j; i \neq j; i, j = 1, \dots, k$, para algum j

Região de rejeição

$$|\bar{R}_i - \bar{R}_j| \geq Z_{\frac{\alpha}{k(k-1)}} \sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

onde $\bar{R}_i$ e $\bar{R}_j$ são respetivamente a média do grupo i e j.

87

Exemplo 4

Os dados da tabela pretendem avaliar o rendimento de culturas divididas por quatro grupos diferentes. Pretende-se avaliar se os dados provém de variáveis igualmente distribuídas.

I	II	III	IV
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91	81	
90	89		
84			

88

Resolução

Tabela de ordenação dos dados

Ord.	X_{ij}	$R(X_{ij})$
1	77	1
2	79	2
3	78	3
4	80	4
5	81	6,5
6	81	
7	81	
8	81	
9	82	9
10	83	11
11	83	
12	83	
13	84	13,5
14	84	
15	88	15
16	89	17
17	89	
18	89	

Ord.	X_{ij}	$R(X_{ij})$
19	90	19,5
20	90	
21	91	23
22	91	
23	91	
24	91	
25	91	
26	92	26
27	93	27
28	94	28,5
29	94	
30	95	30
31	96	31,5
32	96	
33	100	33
34	101	34

I	II	II	IV	$R(X_{i1})$	$R(X_{i2})$	$R(X_{i3})$	$R(X_{i4})$	$R(X_{ij})^2$	$R(X_{ij})^2$	$R(X_{ij})^2$	$R(X_{ij})^2$
83	91	101	78	11	23	34	2	121	529	1156	4
91	90	100	82	23	19,5	33	9	529	380,25	1089	81
94	81	91	81	28,5	6,5	23	6,5	812,25	45,25	529	42,25
89	83	93	77	17	11	27	1	289	121	729	1
89	84	96	79	17	13,5	31,5	3	289	182,25	992,25	9
96	83	95	81	31,5	11	30	6,5	992.25	121	900	42,25
91	88	94	80	23	15	28,5	4	529	225	812,25	16
92	91	81		26	23	6,5		676	529	42,25	
90	89			19,5	17			380.25	289		
84				13,5				182.25			
Total								**12112,25**			
R_i				**210**	**139,5**	**213,5**	**32**				
n_i				**10**	**9**	**8**	**7**				
$\frac{R_i^2}{n_i}$				**4410**	**2162,25**	**5697,78**	**146,29**				
n				**34**							

89

Hipóteses a testar

$H_0: \theta_1 = \theta_2 = \theta_3 = \theta_4$

$H_1: \theta_i \neq \theta_j; i \neq j; i, j = 1, \ldots, 4$, para algum j

Estatística de teste

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{{R_i}^2}{n_i} - \frac{n(n+1)^2}{4}\right)$$

Onde,

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} R(X_{ij})^2 - \frac{n(n+1)^2}{4}\right) = \frac{1}{33}\left((11^2 + 23^2 + \cdots + 6{,}5^2 + 4^2) - \frac{34 \times 35^2}{4}\right)$$

$$= \frac{1}{33}\left((121 + 529 + \cdots + 42{,}25 + 16) - 10142{,}5\right) = \frac{1}{33} \times 3251{,}5 = 98{,}530$$

Assim,

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{{R_i}^2}{n_i} - \frac{n(n+1)^2}{4}\right) = \frac{1}{98{,}53}(4410 + \cdots + 146{,}29 - 10142{,}5) = \frac{1}{98{,}53} \times 2003{,}82 = 20{,}337$$

90

Como $k > 3$ e $n_i \geq 5, \forall_j$ e existem empates entre grupos. Assim, o ponto crítico: $\alpha = 0{,}05 \implies \chi^2_{1-\alpha;k-1} = \chi^2_{0{,}95;3} = 7{,}81$

Decisão: rejeita-se H_0 em favor de H_1 com significância 5%, uma vez que $T_{obs} > \chi^2_{0{,}95;3}$, (20,337 > 7,81). Logo, conclui-se que existem diferença significativas entre os valores centrais (mediana) dos grupos.

91

Como se rejeita H_0, efetuam-se agora comparações múltiplas de medianas.

Hipóteses a testar:

$H_0: \theta_i = \theta_j$

$H_1: \theta_i \neq \theta_j; i \neq j; i, j = 1, \dots, k$

Região de rejeição: $|\bar{R}_i - \bar{R}_j| \geq t_{(n-k;1-1/\alpha)} \underbrace{\sqrt{S^2 \frac{n-1-T}{n-k}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}_{\hat{B}}$.

Tabela de comparação das médias

Comp.	$\bar{R}_i$	$\bar{R}_j$	$\lvert\bar{R}_i - \bar{R}_j\rvert$	$t_{30;0.975}$	B	$t_{30;0.975} \times B$	Diferença significativa
I – II	21	15,5	5,50	2,042	3,931	8,027	Não
I – III	21	26,69	5,69		4,058	8,287	Não
I – IV	21	4,57	16,43		4,216	8,610	Sim
II – III	15,5	26,69	11,19		4,157	8,489	sim
II – IV	15,5	4,57	10,93		4,312	8,804	sim
III – IV	26,69	4,57	22,12		4,428	9,042	sim

92

Exercícios

1. Determinado químico é submetido a tratamento por resinas para eliminar impureza. Em certa experiência usaram-se três tipos de resinas. Após passagem através de resinas foram medidas as concentrações de impurezas com os resultados seguintes:

Resina I	Resina II	Resina III
0,046	0,038	0,031
0,025	0,035	0,042
0,014	0,031	0,020
0,017	0,022	0,018
0,043	0,012	0,039

Teste a hipótese de que não há diferenças entre a eficiência das três resinas supondo que as concentrações de impurezas se distribuem normalmente por cada resina.

Solução: $F_{obs} = 0{,}049$

93

2. Realizou-se uma experiência com vista a examinar o efeito da idade na pulsação cardíaca quando indivíduos eram sujeitos a exercício físico. Foram selecionados aleatoriamente 10 indivíduos do sexo feminino de cada um de quatro grupos de idades. Todos percorreram o mesmo percurso, em condições idênticas durante 12 minutos.

≤ 19	20 - 39	40 - 59	≥ 60
29	24	37	28
33	27	25	29
26	33	22	34
27	31	33	36
39	21	28	21
35	28	26	20
33	24	30	25
29	24	24	24
36	21	27	33
22	32	33	32

Verifique se existe evidência suficiente para indicar diferenças as pulsações em relação aos 4 grupos etários. Se houver diferenças identifique os grupos etários diferentes, supondo a normalidade das 4 populações.

Solução: $F_{obs} = 1{,}32$

94

Bibliografia

1. Mello, F.M., 2014. *Dicionário de Estatística. 673 entradas Índice remissivo em Português e inglês. Edições Sílabo, Lisboa.*
2. Hall, A. Neves, C. e Pereira, A., 2011. *Grande Maratona de Estatística no SPSS*. Escolar Editora, Lisboa.
3. Fonseca, J., 2001. *Estatística Matemática. Vol. 2.* Edições Sílabo, Lisboa.
4. Pedrosa, A. C., & Gama, S. M. A. (2016). *Introdução computacional à probabilidade e estatistica com excel*. (Porto Editora, Ed.) (3a Edição). Porto.
5. Pontos Influentes - Análise de Regressão | Portal Action. (n.d.). Retrieved August 16, 2017, from http://www.portalaction.com.br/analise-de-regressao/343-pontos-influentes

Folha prática 3

Universidade Nacional de Timor Lorosa'e (UNTL)

Faculdade de Ciências Exatas (FCE)

Estatística e Análise de Dados ano letivo 2019, 1º semestre

Regressão e análise de variância

1. Os dados a seguir correspondem à variável renda familiar e gasto com alimentação (em unidades monetárias) para uma amostra de 25 famílias.

Renda Familiar (X)	Gasto com Alimentação (Y)
3	1,5
5	2,0
10	6,0
10	7,0
20	10,0
20	12,0
20	15,0
30	8,0
40	10,0
50	20,0
60	20,0
70	25,0
70	30,0
80	25,0
100	40,0
100	35,0
100	40,0
120	30,0
120	40,0
140	40,0
150	50,0
180	40,0
180	50,0
200	60,0
200	50,0

a) Construa o diagrama de dispersão da variável gasto com alimentação (Y) em função da renda familiar (X) e verificar que existe correlação entre as duas variáveis.
b) Calcule o coeficiente de correlação entre essas variáveis. Que poderá concluir?
c) Obtenha a equação de regressão do gasto com alimentação em função da renda familiar.
d) Qual o significado prático do valor da inclinação da reta de regressão da alínea anterior?
e) Considerando o modelo estimado, faça uma previsão do gasto com alimentação se a renda familiar for 125 u.m.

2. A equação de regressão apresentada resume um estudo da relação entre o uso do fumo e a incidência de cancro pulmonar, relacionando o número de anos que uma pessoa fumou com incidência de cancro pulmonar em cada grupo.

$$\widehat{\mathrm{Y}} = -2 + 1{,}70\mathrm{X} \text{ e } \mathrm{r} = 0{,}60$$

a) Explique o significado das estimativas "-2" e "1,70" na equação da regressão.
b) Qual a taxa de incidência de cancro pulmonar para pessoas que fumam há 20 anos?

3. Num processo de fabrico suspeita-se que o número de artigos defeituosos produzidos por uma máquina (Y) depende de velocidade a que essa máquina trabalha (X). A tabela seguinte contém os dados recolhidos, de registos recentes de uma carta de controlo para o número de artigo defeituosos, por um engenheiro e um gestor industrial responsáveis pelo processo de produção:

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	200	400	300	400	200	300	300	400	200	400	200	300
y_i	28	75	37	53	22	58	40	96	46	52	30	60

.

a) Qual dos dois modelos seguintes de regressão linear simples lhe parece mais adequado à descrição dos dados em causa? Justifique.

i. $y = \beta_0 + \beta_1 x + \varepsilon$

ii. $y = \beta_1 x + \varepsilon$

b) Para o modelo escolhido, admita que a velocidade da máquina está agora selecionada em 150. Obtenha uma estimativa pontual do aumento esperado do número de artigos defeituosos produzidos caso tal velocidade duplique.

4. Após ajustamento de certo modelo linear simples a um conjunto de dados foi possível obter os seguintes resultados. Proceda à análise de resíduos e comentos resultados obtidos em relação à validação dos pressupostos do modelo de regressão linear simples.

Obs.	Y	$\hat{Y}$
1	78,5	77,94
2	74,3	80,31
3	104,3	101,61
4	87,6	81,89
5	95,9	98,46
6	109,2	100,83
7	102,7	113,45
8	72,5	81,89
9	93,1	100,04
10	115,9	94,51
11	83,8	88,99
12	113,3	109,52
13	109,4	111,08

5. Em certa região do País foram cuidadosamente medidas 3 variáveis, sempre à mesma hora do dia, em 10 dias aleatoriamente selecionados: pluviosidade, temperatura, velocidade dos ventos. Os resultados obtidos constam da tabela seguinte:

Temperatura	Velocidade vento	Pluviosidade
16	4	5
10	10	11,5
8	16	14
20	3	3
10	9	10,5
26	2	1,5
12	8	10
12	6	8
13	5	7
9	11	13

a) Ajuste o modelo da forma: $Pluviosidade = \beta_0 + \beta_1 Temp + \beta_2 Veloc.Vento + \varepsilon$.

b) Justificando se a temperatura e a velocidade do vento são variáveis eficazes para predizer a pluviosidade nessa região.

c) Em caso afirmativo, utilize o modelo para calcular a pluviosidade predita nessa região, na mesma altura do dia, sabendo que a velocidade do vento era igual a 8,5, sendo de 25^0 de temperatura.

6. A resposta y é uma função de três variáveis independentes, x_1, x_2, x_3 supostamente através de função de regressão, $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$.
   Sendo que numa determinada experiência se obtiveram os dados seguintes:

x_1	x_2	x_3	Y
-3	5	-1	1
-2	0	1	0
-1	-3	1	0
0	-4	0	1
1	-3	-1	2
2	0	-1	3
3	5	1	3

   a) Ajuste o modelo da forma acima.
   b) Há suficiente evidência nos dados que indica que a variável x_3 contribui para predizer Y?
   c) Construa um I.C. de 95% para o valor esperado de Y, através de $x_1 = 1$, $x_2 = -3$ e $x_3 = -1$.
   d) Construa um intervalo de predição de 95% para o valor esperado de Y, com $x_1 = 1$, $x_2 = -3$ e $x_3 = -1$.

7. Um engenheiro é responsável pela redução de custos num processo de produção tentando manter equilibrada a relação "custo-controlo". Um artigo dispendioso neste processo é representado pela quantidade de água usada em cada mês (em litros) para facilidades de produção. Assim, decidiu-se a investigar o uso de água, relacionando-a com os fatores: temperatura média mensal (^{0}F), quantidade de produção (kg/mês), número de dias de elaboração/mês, número de pessoas na lista de pagamentos por mês e número aleatório de dígitos.

   Após de um procedimento implementados no *R* obtendo-se o *output* seguintes:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = dadosx)

Residuals:
    Min      1Q  Median      3Q     Max
-392.96 -188.05  -74.47  243.03  584.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3605.42723 1369.38616   2.633  0.02329 *
x1            11.67723    7.19625   1.623  0.13294
x2             0.07538    0.02409   3.129  0.00959 **
x3          -130.26149   63.25836  -2.059  0.06395 .
x4             2.56340    2.17573   1.178  0.26358
x5             6.02457    4.04675   1.489  0.16465
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 328.8 on 11 degrees of freedom
Multiple R-squared:  0.6033,   Adjusted R-squared:  0.423
F-statistic: 3.346 on 5 and 11 DF,  p-value: 0.04424
```

a) Comenta, o resultado obtido em relação:

i. O modelo ajustado

ii. Significância dos parâmetros

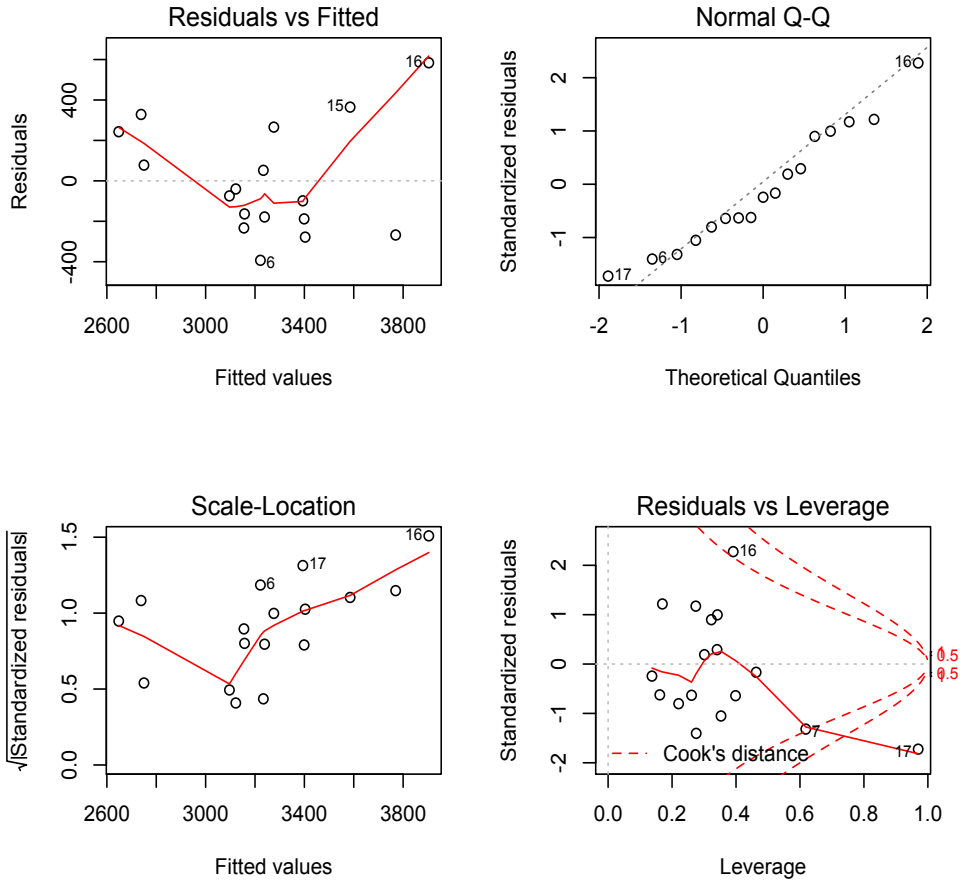
iii. As variâncias dos parâmetros e a variância do erro

iv. Significância do modelo

v. O grau de variação de observação que é explicada pelo modelo

vi. Grau de ajustamento do modelo

b) Considere os gráficos seguinte:



i. Comenta, o que pode concluir sobre análise dos resíduos.

ii. Observa se existe os pontos influentes no modelo.

c) Considere o output do resultado de seleção da variável do método Stepwise seguinte, será que há evidência se todas

```
Start:  AIC=201.64
y ~ x1 + x2 + x3 + x4 + x5

       Df Sum of Sq     RSS    AIC
<none>              1189190 201.65
- x4    1    150066 1339255 201.66
- x5    1    239606 1428796 202.76
- x1    1    284659 1473849 203.29
- x3    1    458411 1647600 205.19
- x2    1   1058640 2247830 210.47

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = dadosx)

Residuals:
    Min      1Q  Median      3Q     Max
-392.96 -188.05  -74.47  243.03  584.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3605.42723 1369.38616   2.633  0.02329 *
x1            11.67723    7.19625   1.623  0.13294
x2             0.07538    0.02409   3.129  0.00959 **
x3          -130.26149   63.25836  -2.059  0.06395 .
x4             2.56340    2.17573   1.178  0.26358
x5             6.02457    4.04675   1.489  0.16465
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 328.8 on 11 degrees of freedom
Multiple R-squared:  0.6033,  Adjusted R-squared:  0.423
F-statistic: 3.346 on 5 and 11 DF,  p-value: 0.04424
```

8. Desejando saber que tipo de filtro se deve usar através do écran de um osciloscópio de raios catódicos, conduziu-se uma experiência para obter registos da intensidade de fixação à qual o observador nota inicialmente o sinal. Os dados constam da tabela seguinte, para 20 repetições da experiência:

Filtro nº.1	Filtro nº.2	Filtro nº.3
90	88	95
87	90	95
93	97	89
96	87	98
94	90	96
88	96	81
90	90	93
84	90	79
101	100	105
96	93	98
90	95	92
82	86	85
93	89	97
90	92	90
96	98	87
87	95	90
99	102	101
101	105	100
79	85	84
98	97	102

a) Teste a igualdade da eficiência dos filtros, supondo a normalidade da intensidade de fixação para cada.

b) Sugira a escolha do filtro mais eficiente. Justifique a sua decisão.

9. Numa secção de uma fábrica têxtil existem cinco teares que fabricam um determinado tipo de tecido. Em princípio, deveriam fabricar a mesma quantidade de tecido por minuto.

   a) Explique qual o plano da experiência que adaptaria para analisar se o pressuposto acima indicado é válido.

   b) O plano anteriormente estabelecido e executado conduziu aos resultados seguintes (peso em gr/min):

Tear				
1	**2**	**3**	**4**	**5**
14,0	14,2	14,4	13,8	13,9
13,9	14,1	14,3	13,6	13,9
14,1	14,2	14,3	13,7	14,0
13,0	14,3	14,5	13,9	13,8
13,8	14,0	14,2	13,6	13,7
14,0	14,4	14,3	13,5	13,8
13,9	14,3	14,4	13,7	13,9

   Teste a hipótese de que os teares fabricam a mesma quantidade de tecido por minuto, admitindo a normalidade dos pesos para cada tear. Se possível, identifique a(s) tear(es) menos produtivo(s).

10. Suspeita-se que 5 livros publicados sob pseudónimos deferentes pertençam ao mesmo autor. Para testar essa hipótese, um perito resolveu analisar aleatoriamente algumas páginas de cada livro no sentido de detetar erros técnicos do mesmo género. Os resultados constam da tabela seguinte.

Livros	**Erros**	
	1	3; 2; 1; 1; 1
	2	2; 2; 1; 2; 2; 4; 0; 1; 2; 2
	3	1; 1; 2; 1; 1; 0
	4	3; 4; 3; 3; 4; 3; 4
	5	2; 2; 1; 3; 2; 2; 2; 1

   a) Supondo que o número de erros desse género, por página, ao longo de um livro tem (aproximadamente) distribuição normal, teste a veracidade da hipótese colocada.

   b) Indique que livros possivelmente pertencerão ao mesmo autor.

11. Três teares foram selecionados aleatoriamente dentre um grande número de teares existentes numa fábrica de têxteis e submetidos a um teste para analisar se eles poderiam ser considerados homogéneos quanto à sua produção. Em intervalos aleatoriamente escolhidos, pesou-se o tecido produzido por cada tear. Após cálculos, obtiveram-se os resultados seguintes:

$$\sum_{i=1}^{12} x_{1i} = 199{,}2\ ; \qquad \sum_{i=1}^{14} x_{2i} = 210\ ; \qquad \sum_{i=1}^{10} x_{3i} = 142;$$

$$\sum_{i=1}^{12} (x_{1i} - \bar{x}_1)^2 = 58{,}8\ ; \qquad \sum_{i=1}^{14} (x_{2i} - \bar{x}_2)^2 = 75{,}15\ ; \qquad \sum_{i=1}^{10} (x_{3i} - \bar{x}_3)^2 = 48{,}8$$

Proceda à análise dos dados na tentativa de testar a hipótese admitida, elaborando a respetiva tabela ANOVA, interprete os resultados obtidos.

12. Um conjunto de 12 provadores atribuiu as seguintes classificações (de 1 a 4) a quatro tipos de vinhos:

	Vinhos			
Provadores	**V1**	**V2**	**V3**	**V4**
1	4	3	2	1
2	4	2	3	1
3	3	1,5	1,5	4
4	3	1	2	4
5	4	2	1	3
6	2	2	2	4
7	1	3	2	4
8	2	4	1	3
9	3,5	1	2	3,5
10	4	1	3	2
11	4	2	3	1
12	3,5	1	2	3,5

Pretende-se saber se as classificações atribuídas aos quatro tipos de vinho são estatisticamente idênticas ($\alpha = 0{,}10$).

Atividades para aprendizagem com R

```
      ############################ CORRELAÇÃO ##############################

x<-c(30,20,60,80,40,50,60,30,70,60)
y<-c(73,50,128,170,87,108,135,69,148,132)
dados<-data.frame(y,x);dados
attach(dados)
plot(y~x,pch=16,main="Diagrama de Dispersão")  # Diagrama de dispersão
r<-cor(x,y);r                                  # coeficiente de correlação

      ############### MODELO DE REGRESSÃO LINEAR SIMPLES ################

modelo<-lm(y~x) # modelo de regressão linear (y~x: modelo y como função
                  estatística de x)
modelo
summary(modelo)

resid(modelo)       # Resíduos
coef(modelo)        # coeficiente da regressão
predict(modelo)     # valores preditos
#fitted(modelo)     # valores ajustados

plot(y~x,pch=16,main="Diagrama de Dispersão")
abline(modelo,lty=1,lwd=2,col="blue")          # reta de regressão (reta de
 tendência)

confint(modelo)                  # I.C. dos parâmetros
#confint(modelo, level=0.95)
anova(modelo)                    # teste de ANOVA (significância do modelo)

## Gráfico para verificar E(e)=0 e var(e)=constante ou homogenidade dos resíduos
plot(predict(modelo),resid(modelo),xlab="Preditos",ylab="Resíduos",pch=16,
 main="Gráfico dos resíduos vs preditos")
abline(h=mean(resid(modelo)),col="blue",lwd=2)
text(150,-0.3,expression(paste('E','',(epsilon[i])==0)),col="blue")

## QQ_plot ==> verificar ei~N(0,miu^2)
qqnorm(resid(modelo),ylab="Resíduos",xlab="Quantis teóricos",main="Normal QQ-
 Plot",pch=16)
qqline(resid(modelo),col="blue",lwd=2)

## Testes de hipóteses
ks.test(resid(modelo),"pnorm",mean(resid(modelo)),sd(resid(modelo))) # teste do
 Kolmogorov-Smirnov
shapiro.test(rstudent(modelo))   # teste de shapiro wilk (normalidade)
lillie.test(resid(modelo))       # Lilliefors
dwtest(modelo)                   # teste de Durbin-Watson (independência)
```

```
## PREVISÃO
x0=data.frame(x=40)
predict(modelo,x0,interval="confidence")       # previsão em média
predict(modelo,x0,interval="prediction")       # previsão pontual
x0=data.frame(x=seq(0,80,1))
p1=predict(modelo,x0,interval="confidence",se=T)
p2=predict(modelo,x0,interval="prediction",se=T)

## Representação gráfica das previsões ##
matplot(x0,p1$fit,lty=c(1,2,2),type="l",lwd=2,xlab="Tamanho de lote",ylab="Horas
 de serviço",col="green",main="Representação dos intervalos de confiança 95% de
 previsão")
matplot(x0,p2$fit,lty=c(1,2,2),type="l",lwd=2,col="red",add=TRUE)
abline(modelo,col="blue",lwd=2)
legend("bottomright",inset=.05,title="I.C. de previsão",c("Prev. em média","Prev.
 pontual","Reta regressão"),fill=c("green","red","blue"))

par(mfrow=c(2,2))
plot(modelo)        # representação gráfica do modelo


      ############### MODELO DE REGRESSÃO LINEAR MÚLTIPLA #################

X<-matrix(c(1,8,4,2,9,-8,11,-10,3,6,8,-6,6,0,13,-13,2,4,7,-2,6,-
 4),nrow=11,ncol=2,byrow=TRUE)
Y<-matrix(c(6,8,1,0,5,3,2,-4,10,-3,5),nrow=11,ncol=1,byrow=TRUE)

da<-data.frame(Y,X);da
attach(da)

cor(da)             # matriz de correlação
pairs(da,pch=16)

modelo<-lm(Y~X)     # modelo de regressão linear múltipla
# OU
modelo1<-lm(Y~X1+X2); modelo1

## PREVISÃO ##
x01=data.frame(X1=5)
x02=data.frame(X2=3)
x0=c(x01,x02)
predict(modelo1,x0,interval="confidence")      # previsão em média
predict(modelo1,x0,interval="prediction")      # previsão pontual

## PONTOS DE INFLUENTES ##
influence.measures(modelo1)      # DFFITS, DFBETAS, Distância COOK

#############################################################################
X<-
 matrix(c(3.34,2.73,1.12,1.91,0.89,3.95,3.80,1.35,2.30,1.20,2.53,2.14,0.87,1.45,0
 .82,2.53,2.12,0.87,1.38,0.69,2.68,2.30,0.92,1.63,0.77,2.78,2.37,1.07,1.56,0.79,2
 .75,2.12,1.02,1.30,0.79,3.24,2.75,1.12,1.96,0.87,2.53,2.17,0.12,1.43,0.74,3.42,2
 .88,1.15,1.96,0.94,3.32,2.81,1.15,1.94,0.97,3.39,2.81,1.22,1.96,0.97,4.74,3.75,1
 .53,2.58,1.66,4.03,3.42,1.33,2.42,1.28,5.02,4.06,1.71,3.06,1.50,4.59,3.98,1.68,2
 .60,1.50,4.46,4.03,1.61,2.78,1.50,4.39,3.65,1.63,2.60,1.61,4.28,4.00,1.84,2.45,1
```

```
 .73,4.46,4.06,1.73,2.75,1.58,5.59,2.74,1.91,3.16,1.84,4.41,4.26,1.61,2.91,1.40),
 nrow=22,ncol=5,byrow=TRUE)
Y<-
 matrix(c(0.90,1.33,0.33,0.37,0.50,0.56,0.47,0.89,0.29,0.96,0.91,0.87,3.93,2.06,3
 .90,2.83,3.46,2.92,2.50,3.57,4.66,3.16),nrow=22,ncol=1,byrow=TRUE)

da<-data.frame(Y,X)
attach(da)

modelo<-lm(Y~X)            # modelo de regressão linear múltipla
# OU
modelo1<-lm(Y~X1+X2+X3+X4+X5)

# instal pacote car
vif(modelo1)               # verificar multicolinearidade

var_selec <- step(modelo1,direction="both",trace=TRUE)
summary(var_selec) # usa método backward, outros : forward, both (stepwise),
 quando existe múlticolinearidade

modelo2<-var_selec;modelo2


            ############### ANÁLISE DE VARIÂNCIAS #################

dados<- read.table("AV_1.txt", head=T,sep="\t")      # importar os dados do
 ficheiro de excel
attach(dados)

## Gráfica de médias ##
médias<-tapply(Nota,Tecnica,mean);médias
plot(as.numeric(Tecnica),Nota,pch=16,xlab="Técnica",ylab="Nota",axes=FALSE)
axis(2,las=1)
x1<-1:nlevels(Tecnica)
axis(1,x1,levels(Tecnica))
lines(médias,col="blue")          # gráfica da linha de pontos medios
points(médias,col="red",pch=16)   # pontos médios

ANOVA<-aov(Nota~Tecnica); ANOVA   # modelo de ANOVA

## comparação múltipla
TukeyHSD(ANOVA)
#scheffe.test(ANOVA)
#scheffe.test(ANOVA,"Tecnica")

## homogeneidade de variâncias ##
# package stats
bartlett.test(Nota~Tecnica,data=dados)
# package car
leveneTest(Nota~Tecnica,data=dados)


#############################################################################
dados<- read.table("AV_4.txt", head=T,sep="\t");attach(dados)
kruskal.test (Rendimento,Grupo)
```

Parte IV Conclusão final

Atualmente, a Estatística é uma ferramenta de grande importância para nossa sociedade, e é correntemente usada no dia-a-dia profissional. A Estatística ajuda a planear a aquisição de dados, a interpretar e a analisar os dados obtidos e a apresentar os resultados de maneira a facilitar a decisão, nas diferentes áreas do conhecimento humano. Ela não se resume apenas a números e a gráficos, mas constitui uma ferramenta essencial que auxilia nas respostas às perguntas formuladas, permitindo uma descrição clara e objetiva dos fenómenos em estudo. O estudo da Estatística também permite assim o desenvolvimento de aptidões, em particular, a organização, o senso critico e análise de resultados.

A Estatística também é importante no contexto da sociedade moderna para a formação do cidadão, especialmente do aluno que fará, num futuro próximo, parte do mundo do trabalho. Para exercer a cidadania, especialmente numa sociedade centrada para o conhecimento, é fundamental que os alunos saibam comunicar ideias, executar procedimentos, construir e interpretar tabelas e gráficos, fazer estimativas e inferências lógicas e analisar dados e informações. Assim sendo, a estatística contribui *significativamente* para o desenvolvimento dessas capacidades e, por estes motivos, a existência de Unidades Curriculares de Estatística e Análise de Dados em oferta formativa no ensino superior é incontornável.

Neste trabalho desenvolveu-se o material pedagógico e de apoio à lecionação da Unidade Curricular "*Estatística e Análise de Dados (EAD)*", uma das novas UC que irão ser lecionadas na Faculdade de Ciências Exatas da Universidade Nacional de Timor Lorosa'e (FCE/UNTL). Na análise exploratória de dados, usamos os métodos como tabela, gráficos e medidas amostrais para tentar entender quais são as estruturas fundamentais dos dados que queremos analisar; e se analisamos dados vindos de fontes ou variáveis diferentes, também tentamos entender as estruturas que relacionam as fontes entre si. Em construção de modelos estatísticos que representam adequadamente relação entre variáveis, neste caso, em particular o modelo log-lineares ordinais (para variáveis qualitativas), modelos regressão linear (para variáveis quantitativas), análise de variância (ANOVA para uma variável dependente quantitativa e uma variável independente quantitativa) e quantificar os graus de relacionamento entre as variáveis estudadas.

É importante ressaltar a abordagem computacional da matéria para tornar a aprendizagem, mais aliciante, pode funcionar como laboratório para simular e experimentação, possibilitando aos estudantes adquirir maior intuição e compreender mais facilmente a matéria. Por outro lado, o computador permite resolver problemas cuja solução pode não ser obtida de forma analítica.

Considera-se este trabalho apresenta algumas limitações em aspectos distintos. Assim, acrescentam-se algumas sugestõs para futuro trabalho e melhoramentos que se consideram relevante para este material pedagógico.

Acrescentar alguns conteúdos relacionados para completar e melhorar o material pedagógico desta Unidade Curricular, por ex.

- análise ANOVA com uma factor para amostras emparelhadas (ou medições repetidas), incluindo abordagens paramétricas e não paramétricas.
- uma introdução ao método de regressão linear robusta, como alternativa à regressão linear tradicional para quando a variável resposta não segue uma distribuição normal ou apresenta valores discrepante

Também irá surgir a necessidade de completar os slides de atividade passo-a-passo no R para orientação tutorial em contexto de aula prática. Neste contexto, irão ser preparados projectos para realização de trabalhos em grupo e apresentação na sala de aula, com o objectivo de desenvolver a capacidade do aluno utilizar em contexto muito prático os conteúdos estudados nesta unidade curricular.

Finalmente, após a primeira edição do curso, irá haver um esperado reajuste na "Distribuição de horário e plano da aula" por forma a adaptar os conteúdos (e os tempos de alocação a cada tema) às dificuldades e às necessidades de aprendizagem dos alunos.

Espera-se que este trabalho estimule mais (e melhor) a aprendizagem na área da estatística e, em concreto, que contribua para o desenvolvimento do material pedagógico e de apoio à lecionação da Unidade Curricular de Estatística e Análise de Dados no ensino superior em Timor Lorosa'e.

Anexos

1. Distribuição de horário e plano da aula

2. Tabelas de Distribuição para uso em aula

2.1. Tabela da Distribuição Qui-Quadrado

2.2. Tabela da Distribuição Normal Reduzida

2.3. Tabela da Distribuição t-Student

2.4. Tabela da Distribuição F – Snedecor a $\alpha = 0{,}05$

2.5. $W_{\alpha,n}$ (os valores críticos da estatística W de Shapiro-Wilk)

2.6. Tabela de coeficiente a_i do teste de Shapiro-Wilk

2.7. Valores críticos de $d_{\alpha,n}$ de Kolmogorov-Smirnov

2.8. Quantis da estatística de Lilliefors para a distribuição Normal

2.9. Tabela de valores críticos de Durbin-Watson $\alpha = 0{,}05$

2.10. Tabela valores críticos da Distribuição t-Studentized Range

2.11. Tabela de quantis da estatística de Kruskal-Wallis para pequenas amostras

Distribuição de horário e plano da aula

Anexo 1

DISTRIBUIÇÃO DO HORÁRIO				
Tópicos	**Subtópicos**	**Horas**	**Total horas**	**Nº. aula**
Análise exploratório de dados	1. Revisão de conceitos de estatística descritiva	2	10	5 aulas
	2. Organização de dados	2		
	3. Medidas amostrais	6		
Análise de tabelas de contingência	1. Revisão de conceito de probabilidade	4	38	19 aulas
	2. Teste de hipóteses	14		
	3. Tabelas contingência r x c	6		
	4. Tabelas contingência r x c x l	14		
Regressão linear e análise de variância (ANOVA)	1. Introdução	2	38	19 aulas
	2. Correlação e regressão			
	3. Regressão linear simples	14		
	4. Regressão linear múltipla	12		
	5. Análise de variância (ANOVA)	10		
Avaliação contínua	1º teste	2	4	2 aulas
	2º teste	2		
Total			**90**	**45 aulas**

**Cada aula equivalente com 02h00, total horas de encontro cada semana é 06h00, assim equivalente com 3 aulas por semana.*

PLANO DE AULA			
Nº. de Aula	**Atividades**	**Metas de aprendizagem**	**Estudo Autónomo**
Aula 1 (ATP)	Apresentação e informações gerais sobre o funcionamento da disciplina; bibliogafia e componentes de avaliação. Introdução a análise exploratória de dados; Revisão conceito de estatística descritivas.	Os alunos conhecem as caraterísticas e sabem como funciona a disciplina. Esperar que os alunos entenderem sobre o conceito de estatística descritiva.	Resolver exercícios e consultar outras referências bibliográficas no sentido de aprofundar os conceitos de estatística descritivas.
Aula 2 (ATP)	Organização de dados numa tabela e sua representação gráfica; exemplos; exercícios.	Entenderem de organizar os dados numa tabela de frequência e ser capaz de representar nos gráficos, tanto para variáveis discretas e variáveis contínuas.	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 3 (ATP)	Medidas amostrais: medidas de localização; medidas de dispersão; exemplos; exercícios.	Sabem calcular as medidas de localização e medidas de dispersão	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 4 (ATP)	Medidas de assimetria (skewness); representação gráfica de auxilio de medidas amostrais (boxplot);	Saber calcular as medidas de assimetria e saber construir o boxplot. Ser capaz de identificar o sinal de assimetria através de comparação das medidas de amostrais e de visualização de boxplot.	Resolver exercícios indicados e consultar outras referências bibliográficas .
Aula 5 (AP)	Resolução de exercícios da aula anterior e da folha prática 1. Apresentação o software computacional R e resolver os exercícios propostos nas aulas anteriores através do R.	Ser capaz de resolver os problemas propostos. Conhecer, entender e saber usar o software R para resolver os exercícios propostos.	Resolver exercícios indicados através do R e, consultar outras referências bibliográficas .
Aula 6 (ATP)	Análise de tabelas de contingência: introdução; revisão de conceito de probabilidade.	Conhecer os aspetos que contem numas tabelas de contingência; relembrar os conceitos de probabilidades, e esperar que os alunos saber usar em conteúdos que vão estudar a seguir.	Consultar outras referências bibliográficas .
Aula 7 (ATP)	Continuação da aula anterior: revisão de conceito de probabilidade; exemplos; resolver os exercícios e folha prática 2.	Saber calcular a probabilidade de um acontecimento. Saber resolver os exercícios sobre as probabilidades.	Resolver exercícios indicados e consultar outras referências bibliográficas .

Aula 8 (ATP)	Teste de hipóteses: Teste de independência de Qui-quadrado e de razão verossimilhança.	Ser capaz em análise de dados de variáveis categóricas e amostra independentes numa tabela de contingência, saber calcular as frequências esperadas.	consultar outras referências bibliográficas sobre teste de independência de Qui-quadrado.
Aula 9 (ATP)	Continuação da aula anterior: Teste de independência de Qui-quadrado e teste de razão verossimilhança; exemplo; exercício.	Saber aplicar o teste de Qui-quadrado e de razão verossimilhança; saber determinar os valores críticos através de uma tabela de distribuição e ser capaz em tirar uma conclusão sobre a independências das variáveis baseado no resultado do teste estatístico.	Resolver exercícios indicados.
Aula 10 (ATP)	Continuação da aula anterior: medidas de associação exemplo; exercício; Aplicar o software R em resolução do exemplo e do exercício.	Saber calcular o grau e o sinal de associação entre as variáveis, no caso hipótese de independência é rejeitada. Saber aplicar o R para resolver os exercícios propostos.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas .
Aula 11 (ATP)	Outros teste de Qui-quadrado: Teste de homogeneidade; Teste de ajustamento; exemplo; exercício.	Saber aplicar o teste de Qui-quadrado para testar a hipótese de homogeneidade e ajustamento de um conjunto de dados.	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 12 (ATP)	Teste alternativa de dependências para tabelas 2 x 2: teste exato de Fisher; exemplos.	Saber aplicar o teste alternativa numa situação quando o teste de Qui-quadrado não se aplicável, para amostras independentes e amostra emparelhados ou correlacionados	Resolver exercícios e consultar outras referências bibliográficas.
Aula 13 (ATP)	Continuação da aula anterior: teste exato de Fisher; exemplos; exercícios; teste McNemar.	Saber aplicar o teste exato de Fisher e teste McNemar.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas.
Aula 14 (AP)	Resolver os problemas do exemplo, exercícios ou folha pratica 2 usando R.	Saber aplicar o R para resolver os exercícios propostos.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas.

Aula 15 (ATP)	Tabelas de contingência r x c : localização de fontes de dependência por analise de resíduos; modelos de log-lineares.	identificar quais as categorias de variáveis que mais contribuem para a dependência entre as variáveis. Conhecer e entender os modelos de log-lineares (independência e saturado) de duas variáveis categorizadas, organizados numa tabela bidimensional	Consultar outras referências bibliográficas.
Aula 16 (ATP)	Continuação da aula anterior: modelos de log-lineares; exemplo.	Saber ajustar o modelo e saber estimar os parâmetros do modelo (ou saber calcular os efeitos das variáveis).	Consultar outras referências bibliográficas.
Aula 17 (AP)	Continuação da aula anterior: aplicar o software R em resolução do exemplo, exercícios ou folha prática 2.	Saber aplicar o R para resolver os exercícios propostos.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas.
Aula 18 (ATP)	Tabelas de contingência r x c x l: hipóteses de independências	Saber e entender sobre os tipos de independências entre as variáveis categorizadas numa tabela de contingência tridimensional. Conhecer as hipóteses de independências das três variáveis entre si (independência mutua), de duas variáveis relativamente ao terceiro (independência parcial), duas variáveis são independentes para uma categoria específica da terceira (independência condicional) e a relação parcial condicional entre quaisquer duas variáveis é a mesma para cada nível da terceira variável (associação 2 a 2).	Consultar outras referências bibliográficas.
Aula 19 (ATP)	Continuação da aula anterior: tabelas de contingência r x c x l: hipóteses de independências; exemplo.	Saber calcular as frequências esperadas para cada hipótese de independência e aplicar nos testes de hipóteses.	Consultar outras referências bibliográficas.
Aula 20 (ATP)	Continuação da aula anterior: exemplo; modelos log-lineares.	Rever e continuar a resolver o problema do exemplo da aula anterior. Conhecer os modelos log-lineares (independências e saturado) numa tabela tridimensional.	Resolver exercícios indicados e consultar outras referências bibliográficas.

Aula 21 (ATP)	Continuação da aula anterior: modelos log-lineares.	Saber fazer seleção de modelos: entender os modelos encaixados e saber fazer comparação dos modelos.	Consultar outras referências bibliográficas.
Aula 22 (ATP)	Continuação da aula anterior: modelos log-lineares.	Entender de selecionar o modelo adequado a ajustar aos dados a partir de processo de seleção dos modelos.	Consultar outras referências bibliográficas.
Aula 23 (ATP)	Continuação da aula anterior: modelos log-lineares; exemplo	Saber de selecionar os modelos adequados e saber estimar os parâmetros do modelo.	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 24 (AP)	Continuação da aula anterior: aplicar o software R em resolução do exemplo e do exercício.	Saber aplicar o R para resolver os exercícios propostos.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas.
Aula 25	Avaliação contínua: 1º teste	Avaliar os conhecimentos e aproveitamentos dos alunos sobre as matérias que tinham de estudar.	
Aula 26 (ATP)	Regressão linear e analise variância (ANOVA): Introdução; correlação e regressão; exemplo.	Entender a diferença de aplicação de análise de regressão e da ANOVA. saber analisar o relacionamento entre as variáveis (através visualização do diagrama de dispersão), medir o grau de relacionamento (através de coeficiente de correlação) e estimar o relacionamento entre as variáveis por meio de uma equação matemática que melhor descreve a relação.	Consultar outras referências bibliográficas e resolver exercícios indicados.
Aula 27 (ATP)	Regressão linear simples: modelo de regressão linear simples; estimação dos parâmetros do modelo; exemplo.	Saber estimar os parâmetros do modelo e descrever a relação linear entre duas variáveis na forma matemática, através da equação $y = \beta_0 + \beta_1 x + \varepsilon_i$.	Consultar outras referências bibliográficas e resolver exercícios indicados.

Aula 28 (ATP)	Continuação da aula anterior, regressão linear simples: inferência sobre os parâmetros do modelo; exemplo.	Conhecer a distribuição amostral dos parâmetros, saber fazer o teste de significância e calcular o intervalo de confiança dos parâmetros do modelo	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 29 (ATP)	Continuação da aula anterior, regressão linear simples: significado, avaliação da qualidade do modelo; exemplo.	Saber verificar o significado estatístico do modelo através de ANOVA da regressão; saber avaliar a qualidade do modelo através de coeficiente de determinação e saber calcular o grau de ajustamento do modelo.	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 30 (ATP)	Continuação da aula anterior, regressão linear simples: validação dos pressupostos do modelo; exemplo.	Saber verificar os pressupostos do modelo através de análise de resíduos ($E[\varepsilon_i\|x] = 0$ *e* $var[\varepsilon_i\|x] = \sigma^2,\ i = 1, \dots, n;\ \varepsilon_i \sim N(0, \sigma^2)$ *e* ε_i *são independentes*), neste caso, através de visualização gráfica e teste de hipóteses.	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 31 (ATP)	Continuação da aula anterior, regressão linear simples: validação dos pressupostos do modelo; exemplo.	Saber aplicar o teste de Kolmogorov-Smirnov (KS), Lilliefors e Shapiro-Wilk para verificar a normalidade dos resíduos (erros); saber aplicar o teste de Durbin-Watson para verificar a independências dos dois resíduos sucessivos.	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 32 (ATP)	Continuação da aula anterior, regressão linear simples: continuação de resoluções do exemplo; previsão em média e pontual; exemplo.	Saber calcular o intervalo de confiança para previsão em media e previsão pontual.	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 33 (AP)	Aplicar o software R em resoluções dos exemplos, exercícios e folha prática 2.	Saber aplicar o R para resolver os exercícios propostos.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas.

Aula 34 (ATP)	Regressão linear múltipla: modelo de regressão linear múltipla; estimação e inferências dos parâmetros do modelo.	Saber estimar os parâmetros do modelo e descrever a relação linear entre uma variável dependentes com dois ou mais variáveis independentes na forma matemática, através da equação $$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$	Consultar outras referências bibliográficas.
Aula 35 (ATP)	Continuação da aula anterior, regressão linear múltipla: significado estatístico do modelo; avaliação da qualidade do modelo; previsão em média e pontual; exemplo.	Saber verificar o significado estatístico do modelo através de ANOVA da regressão; saber avaliar a qualidade do modelo através de coeficiente de determinação e saber calcular o grau de ajustamento do modelo. Saber calcular o intervalo de confiança para previsão em media e previsão pontual.	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 36 (ATP)	Continuação da aula anterior, regressão linear múltipla: resolução do exemplo através do R.	Saber resolver o problema do modelo de regressão linear múltipla.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas.
Aula 37 (ATP)	Continuação da aula anterior, regressão linear múltipla: diagnostico de pontos influentes; multicolinearidade; seleção de variáveis numa regressão múltipla; resolução do exemplo através do R.	Saber identificar os pontos influentes, existência de multicolinearidade e saber fazer a seleção de variáveis através do R.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas.
Aula 38 (ATP)	Continuação da aula anterior, regressão linear múltipla: seleção de variáveis numa regressão múltipla; resolução do exemplo através do R.	Saber ler o output do R.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas.
Aula 39 (AP)	Resoluções dos problemas dos exercícios e da folha prática 3 através do R.	Saber aplicar o R para resolver os exercícios propostos.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas.

Aula 40 (ATP)	ANOVA: ANOVA com um fator e efeitos fixos; teste estatístico; comparações múltiplas; exemplo.	Conhecer o modelo de ANOVA com um fator e efeitos fixos; saber aplicar o teste estatístico; saber identificar as diferenças estatisticamente significativas das médias dos grupos através de comparações de múltiplas, que feita pelo teste de Tukey ou teste de Scheffé.	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 41 (ATP)	Continuação da aula anterior, ANOVA com um fator e efeitos aleatórios; teste estatístico; exemplo.	Conhecer o modelo de ANOVA com um fator e efeitos fixos; saber aplicar o teste estatístico; saber estimar a variância do fator no caso rejeitar a hipótese nula.	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 42 (ATP)	Continuação da aula anterior, avaliar os pressupostos da ANOVA a um fator; resolução do exemplo através do R.	Saber testar a homogeneidade de variâncias através de teste de Bartlett ou teste de Levene.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas.
Aula 43 (ATP)	Continuação da aula anterior, ANOVA não paramétrica: teste de Kruskal-Wallis; exemplo.	Saber aplicar o teste alternativa para ANOVA com um fator quando os pressupostos de normalidade e homogeneidade são violados.	Resolver exercícios indicados e consultar outras referências bibliográficas.
Aula 44 (AP)	Resoluções dos problemas dos exercícios e da folha prática 3 através do R.	Saber aplicar o R para resolver os exercícios propostos.	Resolver exercícios indicados usando R e consultar outras referências bibliográficas.
Aula 45	Avaliação contínua: 2º teste	Avaliar os conhecimentos e aproveitamentos dos alunos sobre as restantes matérias que ainda não incluem no 1º teste.	

Tabelas de Distribuição para uso em aula

Anexo 2.1.

Tabela da Distribuição Qui-Quadrado

	$1-\alpha$													
n	0,005	0,01	0,025	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,975	0,99	0,995	
1	3,93E-05	0,000157	0,000982	0,003932	0,016	0,102	0,455	1,323	2,706	3,841	5,024	6,635	7,879	1
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,773	4,605	5,991	7,378	9,210	10,597	2
3	0,072	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	9,348	11,345	12,838	3
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	9,488	11,143	13,277	14,860	4
5	0,412	0,554	0,831	1,145	1,610	2,675	4,351	6,626	9,236	11,070	12,832	15,086	16,750	5
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	14,449	16,812	18,548	6
7	0,989	1,239	1,690	2,167	2,833	4,255	6,346	9,037	12,017	14,067	16,013	18,475	20,278	7
8	1,344	1,647	2,180	2,733	3,490	5,071	7,344	10,219	13,362	15,507	17,535	20,090	21,955	8
9	1,735	2,088	2,700	3,325	4,168	5,899	8,343	11,389	14,684	16,919	19,023	21,666	23,589	9
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342	12,549	15,987	18,307	20,483	23,209	25,188	10
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	21,920	24,725	26,757	11
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340	14,845	18,549	21,026	23,337	26,217	28,300	12
13	3,565	4,107	5,009	5,892	7,041	9,299	12,340	15,984	19,812	22,362	24,736	27,688	29,819	13
14	4,075	4,660	5,629	6,571	7,790	10,165	13,339	17,117	21,064	23,685	26,119	29,141	31,319	14
15	4,601	5,229	6,262	7,261	8,547	11,037	14,339	18,245	22,307	24,996	27,488	30,578	32,801	15
16	5,142	5,812	6,908	7,962	9,312	11,912	15,338	19,369	23,542	26,296	28,845	32,000	34,267	16
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338	20,489	24,769	27,587	30,191	33,409	35,718	17
18	6,265	7,015	8,231	9,390	10,865	13,675	17,338	21,605	25,989	28,869	31,526	34,805	37,156	18
19	6,844	7,633	8,907	10,117	11,651	14,562	18,338	22,718	27,204	30,144	32,852	36,191	38,582	19
20	7,434	8,260	9,591	10,851	12,443	15,452	19,337	23,828	28,412	31,410	34,170	37,566	39,997	20
21	8,034	8,897	10,283	11,591	13,240	16,344	20,337	24,935	29,615	32,671	35,479	38,932	41,401	21
22	8,643	9,542	10,982	12,338	14,041	17,240	21,337	26,039	30,813	33,924	36,781	40,289	42,796	22
23	9,260	10,196	11,689	13,091	14,848	18,137	22,337	27,141	32,007	35,172	38,076	41,638	44,181	23
24	9,886	10,856	12,401	13,848	15,659	19,037	23,337	28,241	33,196	36,415	39,364	42,980	45,558	24
25	10,520	11,524	13,120	14,611	16,473	19,939	24,337	29,339	34,382	37,652	40,646	44,314	46,928	25
26	11,160	12,198	13,844	15,379	17,292	20,843	25,336	30,435	35,563	38,885	41,923	45,642	48,290	26
27	11,808	12,878	14,573	16,151	18,114	21,749	26,336	31,528	36,741	40,113	43,195	46,963	49,645	27
28	12,461	13,565	15,308	16,928	18,939	22,657	27,336	32,620	37,916	41,337	44,461	48,278	50,994	28
29	13,121	14,256	16,047	17,708	19,768	23,567	28,336	33,711	39,087	42,557	45,722	49,588	52,335	29
30	13,787	14,953	16,791	18,493	20,599	24,478	29,336	34,800	40,256	43,773	46,979	50,892	53,672	30
40	20,707	22,164	24,433	26,509	29,051	33,660	39,335	45,616	51,805	55,758	59,342	63,691	66,766	40
50	27,991	29,707	32,357	34,764	37,689	42,942	49,335	56,334	63,167	67,505	71,420	76,154	79,490	50
60	35,534	37,485	40,482	43,188	46,459	52,294	59,335	66,981	74,397	79,082	83,298	88,379	91,952	60
70	43,275	45,442	48,758	51,739	55,329	61,698	69,334	77,577	85,527	90,531	95,023	100,425	104,215	70
80	51,172	53,540	57,153	60,391	64,278	71,145	79,334	88,130	96,578	101,879	106,629	112,329	116,321	80
90	59,196	61,754	65,647	69,126	73,291	80,625	89,334	98,650	107,565	113,145	118,136	124,116	128,299	90
100	67,328	70,065	74,222	77,929	82,358	90,133	99,334	109,141	118,498	124,342	129,561	135,807	140,170	100

Fonte: Fonseca, J. (2001). *Estatística Matemática*. (Edições Sílaba, Ed.) (1a Edição). Lisboa.

Anexo 2.2.

Tabela da Distribuição Normal Reduzida

z	Φ(-z)	Φ(z)
0.01	0.4960	0.5040
0.02	0.4920	0.5080
0.03	0.4880	0.5120
0.04	0.4840	0.5160
0.05	0.4801	0.5199
0.06	0.4761	0.5239
0.07	0.4721	0.5279
0.08	0.4681	0.5319
0.09	0.4641	0.5359
0.10	0.4602	0.5398
0.11	0.4562	0.5438
0.12	0.4522	0.5478
0.13	0.4483	0.5517
0.14	0.4443	0.5557
0.15	0.4404	0.5596
0.16	0.4364	0.5636
0.17	0.4325	0.5675
0.18	0.4286	0.5714
0.19	0.4247	0.5753
0.20	0.4207	0.5793
0.21	0.4168	0.5832
0.22	0.4129	0.5871
0.23	0.4090	0.5910
0.24	0.4052	0.5948
0.25	0.4013	0.5987
0.26	0.3974	0.6026
0.27	0.3936	0.6064
0.28	0.3897	0.6103
0.29	0.3859	0.6141
0.30	0.3821	0.6179
0.31	0.3783	0.6217
0.32	0.3745	0.6255
0.33	0.3707	0.6293
0.34	0.3669	0.6331
0.35	0.3632	0.6368
0.36	0.3594	0.6406
0.37	0.3557	0.6443
0.38	0.3520	0.6480
0.39	0.3483	0.6517
0.40	0.3446	0.6554
0.41	0.3409	0.6591
0.42	0.3372	0.6628
0.43	0.3336	0.6664
0.44	0.3300	0.6700
0.45	0.3264	0.6736
0.46	0.3228	0.6772
0.47	0.3192	0.6808
0.48	0.3156	0.6844
0.49	0.3121	0.6879
0.50	0.3085	0.6915

z	Φ(-z)	Φ(z)
0.51	0.3050	0.6950
0.52	0.3015	0.6985
0.53	0.2981	0.7019
0.54	0.2946	0.7054
0.55	0.2912	0.7088
0.56	0.2877	0.7123
0.57	0.2843	0.7157
0.58	0.2810	0.7190
0.59	0.2776	0.7224
0.60	0.2743	0.7257
0.61	0.2709	0.7291
0.62	0.2676	0.7324
0.63	0.2643	0.7357
0.64	0.2611	0.7389
0.65	0.2578	0.7422
0.66	0.2546	0.7454
0.67	0.2514	0.7486
0.68	0.2483	0.7517
0.69	0.2451	0.7549
0.70	0.2420	0.7580
0.71	0.2389	0.7611
0.72	0.2358	0.7642
0.73	0.2327	0.7673
0.74	0.2296	0.7704
0.75	0.2266	0.7734
0.76	0.2236	0.7764
0.77	0.2206	0.7794
0.78	0.2177	0.7823
0.79	0.2148	0.7852
0.80	0.2119	0.7881
0.81	0.2090	0.7910
0.82	0.2061	0.7939
0.83	0.2033	0.7967
0.84	0.2005	0.7995
0.85	0.1977	0.8023
0.86	0.1949	0.8051
0.87	0.1922	0.8078
0.88	0.1894	0.8106
0.89	0.1867	0.8133
0.90	0.1841	0.8159
0.91	0.1814	0.8186
0.92	0.1788	0.8212
0.93	0.1762	0.8238
0.94	0.1736	0.8264
0.95	0.1711	0.8289
0.96	0.1685	0.8315
0.97	0.1660	0.8340
0.98	0.1635	0.8365
0.99	0.1611	0.8389
1.00	0.1587	0.8413

z	Φ(-z)	Φ(z)
1.01	0.1562	0.8438
1.02	0.1539	0.8461
1.03	0.1515	0.8485
1.04	0.1492	0.8508
1.05	0.1469	0.8531
1.06	0.1446	0.8554
1.07	0.1423	0.8577
1.08	0.1401	0.8599
1.09	0.1379	0.8621
1.10	0.1357	0.8643
1.11	0.1335	0.8665
1.12	0.1314	0.8686
1.13	0.1292	0.8708
1.14	0.1271	0.8729
1.15	0.1251	0.8749
1.16	0.1230	0.8770
1.17	0.1210	0.8790
1.18	0.1190	0.8810
1.19	0.1170	0.8830
1.20	0.1151	0.8849
1.21	0.1131	0.8869
1.22	0.1112	0.8888
1.23	0.1093	0.8907
1.24	0.1075	0.8925
1.25	0.1056	0.8944
1.26	0.1038	0.8962
1.27	0.1020	0.8980
1.28	0.1003	0.8997
1.29	0.0985	0.9015
1.30	0.0968	0.9032
1.31	0.0951	0.9049
1.32	0.0934	0.9066
1.33	0.0918	0.9082
1.34	0.0901	0.9099
1.35	0.0885	0.9115
1.36	0.0869	0.9131
1.37	0.0853	0.9147
1.38	0.0838	0.9162
1.39	0.0823	0.9177
1.40	0.0808	0.9192
1.41	0.0793	0.9207
1.42	0.0778	0.9222
1.43	0.0764	0.9236
1.44	0.0749	0.9251
1.45	0.0735	0.9265
1.46	0.0721	0.9279
1.47	0.0708	0.9292
1.48	0.0694	0.9306
1.49	0.0681	0.9319
1.50	0.0668	0.9332

z	Φ(-z)	Φ(z)
1.51	0.0655	0.9345
1.52	0.0643	0.9357
1.53	0.0630	0.9370
1.54	0.0618	0.9382
1.55	0.0606	0.9394
1.56	0.0594	0.9406
1.57	0.0582	0.9418
1.58	0.0571	0.9429
1.59	0.0559	0.9441
1.60	0.0548	0.9452
1.61	0.0537	0.9463
1.62	0.0526	0.9474
1.63	0.0516	0.9484
1.64	0.0505	0.9495
1.65	0.0495	0.9505
1.66	0.0485	0.9515
1.67	0.0475	0.9525
1.68	0.0465	0.9535
1.69	0.0455	0.9545
1.70	0.0446	0.9554
1.71	0.0436	0.9564
1.72	0.0427	0.9573
1.73	0.0418	0.9582
1.74	0.0409	0.9591
1.75	0.0401	0.9599
1.76	0.0392	0.9608
1.77	0.0384	0.9616
1.78	0.0375	0.9625
1.79	0.0367	0.9633
1.80	0.0359	0.9641
1.81	0.0351	0.9649
1.82	0.0344	0.9656
1.83	0.0336	0.9664
1.84	0.0329	0.9671
1.85	0.0322	0.9678
1.86	0.0314	0.9686
1.87	0.0307	0.9693
1.88	0.0301	0.9699
1.89	0.0294	0.9706
1.90	0.0287	0.9713
1.91	0.0281	0.9719
1.92	0.0274	0.9726
1.93	0.0268	0.9732
1.94	0.0262	0.9738
1.95	0.0256	0.9744
1.96	0.0250	0.9750
1.97	0.0244	0.9756
1.98	0.0239	0.9761
1.99	0.0233	0.9767
2.00	0.0228	0.9772

z	Φ(-z)	Φ(z)
2.01	0.0222	0.9778
2.02	0.0217	0.9783
2.03	0.0212	0.9788
2.04	0.0207	0.9793
2.05	0.0202	0.9798
2.06	0.0197	0.9803
2.07	0.0192	0.9808
2.08	0.0188	0.9812
2.09	0.0183	0.9817
2.10	0.0179	0.9821
2.11	0.0174	0.9826
2.12	0.0170	0.9830
2.13	0.0166	0.9834
2.14	0.0162	0.9838
2.15	0.0158	0.9842
2.16	0.0154	0.9846
2.17	0.0150	0.9850
2.18	0.0146	0.9854
2.19	0.0143	0.9857
2.20	0.0139	0.9861
2.21	0.0136	0.9864
2.22	0.0132	0.9868
2.23	0.0129	0.9871
2.24	0.0125	0.9875
2.25	0.0122	0.9878
2.26	0.0119	0.9881
2.27	0.0116	0.9884
2.28	0.0113	0.9887
2.29	0.0110	0.9890
2.30	0.0107	0.9893
2.31	0.0104	0.9896
2.32	0.0102	0.9898
2.33	0.0099	0.9901
2.34	0.0096	0.9904
2.35	0.0094	0.9906
2.36	0.0091	0.9909
2.37	0.0089	0.9911
2.38	0.0087	0.9913
2.39	0.0084	0.9916
2.40	0.0082	0.9918
2.41	0.0080	0.9920
2.42	0.0078	0.9922
2.43	0.0075	0.9925
2.44	0.0073	0.9927
2.45	0.0071	0.9929
2.46	0.0069	0.9931
2.47	0.0068	0.9932
2.48	0.0066	0.9934
2.49	0.0064	0.9936
2.50	0.0062	0.9938

z	Φ(-z)	Φ(z)
2.51	0.0060	0.9940
2.52	0.0059	0.9941
2.53	0.0057	0.9943
2.54	0.0055	0.9945
2.55	0.0054	0.9946
2.56	0.0052	0.9948
2.57	0.0051	0.9949
2.58	0.0049	0.9951
2.59	0.0048	0.9952
2.60	0.0047	0.9953
2.61	0.0045	0.9955
2.62	0.0044	0.9956
2.63	0.0043	0.9957
2.64	0.0041	0.9959
2.65	0.0040	0.9960
2.66	0.0039	0.9961
2.67	0.0038	0.9962
2.68	0.0037	0.9963
2.69	0.0036	0.9964
2.70	0.0035	0.9965
2.71	0.0034	0.9966
2.72	0.0033	0.9967
2.73	0.0032	0.9968
2.74	0.0031	0.9969
2.75	0.0030	0.9970
2.76	0.0029	0.9971
2.77	0.0028	0.9972
2.78	0.0027	0.9973
2.79	0.0026	0.9974
2.80	0.0026	0.9974
2.81	0.0025	0.9975
2.82	0.0024	0.9976
2.83	0.0023	0.9977
2.84	0.0023	0.9977
2.85	0.0022	0.9978
2.86	0.0021	0.9979
2.87	0.0021	0.9979
2.88	0.0020	0.9980
2.89	0.0019	0.9981
2.90	0.0019	0.9981
2.91	0.0018	0.9982
2.92	0.0018	0.9982
2.93	0.0017	0.9983
2.94	0.0016	0.9984
2.95	0.0016	0.9984
2.96	0.0015	0.9985
2.97	0.0015	0.9985
2.98	0.0014	0.9986
2.99	0.0014	0.9986
3.00	0.0013	0.9987

Fonte: https://www.fep.up.pt/disciplinas/lge302/Outros_Docs/Tabela_Normal.pdf.

Retirado em 6 de Setembro de 2017.

Anexo 2.3.

Tabela da Distribuição t-Student

	$1-\alpha$							
n	0,600	0,750	0,900	0,950	0,975	0,990	0,995	0,9995
1	0,325	1,000	3,078	6,314	12,706	31,821	63,657	636,619
2	0,289	0,816	1,886	2,920	4,303	6,965	9,925	31,598
3	0,277	0,765	1,638	2,353	3,182	4,541	5,841	12,924
4	0,271	0,741	1,533	2,132	2,776	3,747	4,604	8,610
5	0,267	0,727	1,476	2,015	2,571	3,365	4,032	6,869
6	0,265	0,718	1,440	1,943	2,447	3,143	3,707	5,959
7	0,263	0,711	1,415	1,895	2,365	2,998	3,499	5,408
8	0,262	0,706	1,397	1,860	2,306	2,896	3,355	5,041
9	0,261	0,703	1,383	1,833	2,262	2,821	3,250	4,781
10	0,260	0,700	1,372	1,812	2,228	2,764	3,169	4,587
11	0,260	0,697	1,363	1,796	2,201	2,718	3,106	4,437
12	0,259	0,695	1,356	1,782	2,179	2,681	3,055	4,318
13	0,259	0,694	1,350	1,771	2,160	2,650	3,012	4,221
14	0,258	0,692	1,345	1,761	2,145	2,624	2,977	4,140
15	0,258	0,691	1,341	1,753	2,131	2,602	2,947	4,073
16	0,258	0,690	1,337	1,746	2,120	2,583	2,921	4,015
17	0,257	0,689	1,333	1,740	2,110	2,567	2,898	3,965
18	0,257	0,688	1,330	1,734	2,101	2,552	2,878	3,922
19	0,257	0,688	1,328	1,729	2,093	2,539	2,861	3,883
20	0,257	0,687	1,325	1,725	2,086	2,528	2,845	3,850
21	0,257	0,686	1,323	1,721	2,080	2,518	2,831	3,819
22	0,256	0,686	1,321	1,717	2,074	2,508	2,819	3,792
23	0,256	0,685	1,319	1,714	2,069	2,500	2,807	3,768
24	0,256	0,685	1,318	1,711	2,064	2,492	2,797	3,745
25	0,256	0,684	1,316	1,708	2,060	2,485	2,787	3,725
26	0,256	0,684	1,315	1,706	2,056	2,479	2,779	3,707
27	0,256	0,684	1,314	1,703	2,052	2,473	2,771	3,689
28	0,256	0,683	1,313	1,701	2,048	2,467	2,763	3,674
29	0,256	0,683	1,311	1,699	2,045	2,462	2,756	3,660
30	0,256	0,683	1,310	1,697	2,042	2,457	2,750	3,646
40	0,255	0,681	1,303	1,684	2,021	2,423	2,704	3,551
60	0,254	0,679	1,296	1,671	2,000	2,390	2,660	3,460
120	0,254	0,677	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,253	0,674	1,282	1,645	1,960	2,326	2,576	3,291

Fonte: Fonseca, J. (2001). *Estatística Matemática*. (Edições Sílaba, Ed.) (1a Edição). Lisboa.

Anexo 2.4.

Tabela da Distribuição F – Snedcor a $\alpha = 0,05$

m / n	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	243,90	245,95	248,02	249,05	250,10	251,14	252,20	253,25
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35

**__m__ é graus de liberdade do numerador; __n__ é graus de liberdade do denominador*

Fonte**:** TABELA VII Distribuição do F de Snedcor. (n.d.). Retirado em 6 de Setembro de 2017, de http://www.estgv.ipv.pt/PaginasPessoais/psarabando/Estat%C3%ADstica%20%20CA%202009-2010/Tabelas/TabelaFSnedcor.pdf

Anexo 2.5.

$W_{\alpha,n}$ (os valores críticos da estatística W de Shapiro-Wilk)

	Nível de significância								
N	**0,01**	**0,02**	**0,05**	**0,1**	**0,5**	**0,9**	**0,95**	**0,98**	**0,99**
3	0,753	0,756	0,767	0,789	0,959	0,998	0,999	1,000	1,000
4	0,687	0,707	0,748	0,792	0,935	0,987	0,992	0,996	0,997
5	0,686	0,715	0,762	0,806	0,927	0,979	0,986	0,991	0,993
6	0,713	0,743	0,788	0,826	0,927	0,974	0,981	0,986	0,989
7	0,730	0,760	0,803	0,838	0,928	0,972	0,979	0,985	0,988
8	0,749	0,778	0,818	0,851	0,932	0,972	0,978	0,984	0,987
9	0,764	0,791	0,829	0,859	0,935	0,972	0,978	0,984	0,986
10	0,781	0,806	0,842	0,869	0,938	0,972	0,978	0,983	0,986
11	0,792	0,817	0,850	0,876	0,940	0,973	0,979	0,984	0,986
12	0,805	0,828	0,859	0,883	0,943	0,973	0,979	0,984	0,986
13	0,814	0,837	0,866	0,889	0,945	0,974	0,979	0,984	0,986
14	0,825	0,846	0,874	0,895	0,947	0,975	0,980	0,984	0,986
15	0,835	0,855	0,881	0,901	0,950	0,975	0,980	0,984	0,987
16	0,844	0,863	0,887	0,906	0,952	0,976	0,981	0,985	0,987
17	0,851	0,869	0,892	0,910	0,954	0,977	0,981	0,985	0,987
18	0,858	0,874	0,897	0,914	0,956	0,978	0,982	0,986	0,988
19	0,863	0,879	0,901	0,917	0,957	0,978	0,982	0,986	0,988
20	0,868	0,884	0,905	0,920	0,959	0,979	0,983	0,986	0,988
21	0,873	0,888	0,908	0,923	0,960	0,980	0,983	0,987	0,989
22	0,878	0,892	0,911	0,926	0,961	0,980	0,984	0,987	0,989
23	0,881	0,895	0,914	0,928	0,962	0,981	0,984	0,987	0,989
24	0,884	0,898	0,916	0,930	0,963	0,981	0,984	0,987	0,989
25	0,888	0,901	0,918	0,931	0,964	0,981	0,985	0,988	0,989
26	0,891	0,904	0,920	0,933	0,965	0,982	0,985	0,988	0,989
27	0,894	0,906	0,923	0,935	0,965	0,982	0,985	0,988	0,990
28	0,896	0,908	0,924	0,936	0,966	0,982	0,985	0,988	0,990

$W_{\alpha,n}$ **(continuação)**

29	0,898	0,910	0,926	0,937	0,966	0,982	0,985	0,988	0,990
30	0,900	0,912	0,927	0,939	0,967	0,983	0,985	0,988	0,990
31	0,902	0,914	0,929	0,940	0,967	0,983	0,986	0,988	0,990
32	0,904	0,915	0,930	0,941	0,968	0,983	0,986	0,988	0,990
33	0,906	0,917	0,931	0,942	0,968	0,983	0,986	0,989	0,990
34	0,908	0,919	0,933	0,943	0,969	0,983	0,986	0,989	0,990
35	0,910	0,920	0,934	0,944	0,969	0,984	0,986	0,989	0,990
36	0,912	0,922	0,935	0,945	0,970	0,984	0,986	0,989	0,990
37	0,914	0,924	0,936	0,946	0,970	0,984	0,987	0,989	0,990
38	0,916	0,925	0,938	0,947	0,971	0,984	0,987	0,989	0,990
39	0,917	0,927	0,939	0,948	0,971	0,984	0,987	0,989	0,991
40	0,919	0,928	0,940	0,949	0,972	0,985	0,987	0,989	0,991
41	0,920	0,929	0,941	0,950	0,972	0,985	0,987	0,989	0,991
42	0,922	0,930	0,942	0,951	0,972	0,985	0,987	0,989	0,991
43	0,923	0,932	0,943	0,951	0,973	0,985	0,987	0,990	0,991
44	0,924	0,933	0,944	0,952	0,973	0,985	0,987	0,990	0,991
45	0,926	0,934	0,945	0,953	0,973	0,985	0,988	0,990	0,991
46	0,927	0,935	0,945	0,953	0,974	0,985	0,988	0,990	0,991
47	0,928	0,936	0,946	0,954	0,974	0,985	0,988	0,990	0,991
48	0,929	0,937	0,947	0,954	0,974	0,985	0,988	0,990	0,991
49	0,929	0,938	0,947	0,955	0,974	0,985	0,988	0,990	0,991
50	0,930	0,939	0,947	0,955	0,974	0,985	0,988	0,990	0,991

Fonte: Teste de Shapiro-Wilk - Inferência | Portal Action. (n.d.). Retirado em 6 de Setembro de 2017, de http://www.portalaction.com.br/inferencia/64-teste-de-shapiro-wilk

Anexo 2.6.

Tabela de coeficiente a_i do teste de Shapiro-Wilk

i\n	2	3	4	5	6	7	8	9	10	11	12	13	
1	0,7071	0,7071	0,6872	0,6646	0,6431	0,6233	0,6062	0,5888	0,5739	0,5601	0,5475	0,5359	
2			0,1677	0,2413	0,2806	0,3031	0,3164	0,3244	0,3291	0,3315	0,3325	0,3325	
3					0,0875	0,1401	0,1743	0,1976	0,2141	0,2260	0,2347	0,2412	
4							0,0561	0,0947	0,1224	0,1429	0,1586	0,1707	
5									0,0399	0,0695	0,0922	0,1099	
6											0,0303	0,0539	
i\n	14	15	16	17	18	19	20	21	22	23	24	25	
1	0,5251	0,5150	0,5056	0,4968	0,4886	0,4808	0,4734	0,4643	0,4590	0,4542	0,4493	0,4450	
2	0,3318	0,3306	0,3290	0,3273	0,3253	0,3232	0,3211	0,3185	0,3156	0,3126	0,3098	0,3069	
3	0,2460	0,2495	0,2521	0,2540	0,2553	0,2561	0,2565	0,2578	0,2571	0,2563	0,2554	0,2543	
4	0,1802	0,1878	0,1939	0,1988	0,2027	0,2059	0,2085	0,2119	0,2131	0,2139	0,2145	0,2148	
5	0,1240	0,1353	0,1447	0,1524	0,1587	0,1641	0,1686	0,1736	0,1764	0,1787	0,1807	0,1822	
6	0,0727	0,0880	0,1005	0,1109	0,1197	0,1271	0,1334	0,1399	0,1443	0,1480	0,1512	0,1539	
7	0,0240	0,0433	0,0593	0,0725	0,0837	0,0932	0,1013	0,1092	0,115	0,1201	0,1245	0,1283	
8			0,0196	0,0359	0,0496	0,0612	0,0711	0,0804	0,0878	0,0941	0,0997	0,1046	
9					0,0163	0,0303	0,0422	0,0530	0,0618	0,0696	0,0764	0,0823	
10							0,0140	0,0263	0,0368	0,0459	0,0539	0,061	
11									0,0122	0,0228	0,0321	0,0403	
12											0,0107	0,0200	
13												0,0000	
i\n	26	27	28	29	30	31	32	33	34	35	36	37	
1	0,4407	0,4366	0,4328	0,4291	0,4254	0,4220	0,4188	0,4156	0,4127	0,4096	0,4068	0,4040	
2	0,3043	0,3018	0,2992	0,2968	0,2944	0,2921	0,2898	0,2876	0,2854	0,2834	0,2813	0,2794	
3	0,2533	0,2522	0,2510	0,2499	0,2487	0,2475	0,2463	0,2451	0,2439	0,2427	0,2415	0,2403	
4	0,2151	0,2152	0,2151	0,2150	0,2148	0,2145	0,2141	0,2137	0,2132	0,1227	0,2121	0,2116	
5	0,1836	0,1848	0,1857	0,1864	0,1870	0,1874	0,1878	0,1880	0,1882	0,1883	0,1883	0,1883	
6	0,1563	0,1584	0,1601	0,1616	0,1630	0,1641	0,1651	0,1660	0,1667	0,1673	0,1678	0,1683	
7	0,1316	0,1346	0,1372	0,1395	0,1415	0,1433	0,1449	0,1463	0,1475	0,1487	0,1496	0,1505	
8	0,1089	0,1128	0,1162	0,1192	0,1219	0,1243	0,1265	0,1284	0,1301	0,1317	0,1331	0,1344	
9	0,0876	0,0923	0,0965	0,1002	0,1036	0,1066	0,1093	0,1118	0,1140	0,1160	0,1179	0,1196	
10	0,0672	0,0728	0,0778	0,0822	0,0862	0,0899	0,0931	0,0961	0,0988	0,1013	0,1036	0,1056	
11	0,0476	0,0540	0,0598	0,065	0,0697	0,0739	0,0777	0,0812	0,0844	0,0873	0,0900	0,0924	
12	0,0284	0,0358	0,0424	0,0483	0,0537	0,0585	0,0629	0,0669	0,0706	0,0739	0,0770	0,0798	
13	0,0094	0,0178	0,0253	0,032	0,0381	0,0435	0,0485	0,0530	0,0572	0,0610	0,0645	0,0677	
14		0,0000	0,0084	0,0159	0,0227	0,0289	0,0344	0,0395	0,0441	0,0484	0,0523	0,0559	
15				0	0,0076	0,0144	0,0206	0,0262	0,0314	0,0361	0,0404	0,0444	
16						0,0000	0,0068	0,0131	0,0187	0,0239	0,0287	0,0331	
17								0,0000	0,0062	0,0119	0,0172	0,0220	
18										0,0000	0,0057	0,0110	
19												0,0000	

Coeficiente a_i (continuação)

i\n	38	39	40	41	42	43	44	45	46	47	48	49	50
1	0,4015	0,3989	0,3964	0,3940	0,3917	0,3894	0,3872	0,3850	0,3830	0,3808	0,3789	0,3770	0,3751
2	0,2774	0,2755	0,2737	0,2719	0,2701	0,2684	0,2667	0,2651	0,2635	0,2620	0,2604	0,2589	0,2574
3	0,2391	0,2380	0,2368	0,2357	0,2345	0,2334	0,2323	0,2313	0,2302	0,2291	0,2281	0,2271	0,2260
4	0,2110	0,2104	0,2098	0,2091	0,2085	0,2078	0,2072	0,2065	0,2058	0,2052	0,2045	0,2038	0,2032
5	0,1881	0,1880	0,1878	0,1876	0,1874	0,1871	0,1868	0,1865	0,1862	0,1859	0,1855	0,1851	0,1847
6	0,1686	0,1689	0,1691	0,1693	0,1694	0,1695	0,1695	0,1695	0,1695	0,1695	0,1693	0,1692	0,1691
7	0,1513	0,1520	0,1526	0,1531	0,1535	0,1539	0,1542	0,1545	0,1548	0,1550	0,1551	0,1553	0,1554
8	0,1356	0,1366	0,1376	0,1384	0,1392	0,1398	0,1405	0,1410	0,1415	0,1420	0,1423	0,1427	0,1430
9	0,1211	0,1225	0,1237	0,1249	0,1259	0,1269	0,1278	0,1286	0,1293	0,1300	0,1306	0,1312	0,1317
10	0,1075	0,1092	0,1108	0,1123	0,1136	0,1149	0,1160	0,1170	0,1180	0,1189	0,1197	0,1205	0,1212
11	0,0947	0,0967	0,0986	0,1004	0,1020	0,1035	0,1049	0,1062	0,1073	0,1085	0,1095	0,1105	0,1113
12	0,0824	0,0848	0,0870	0,0891	0,0909	0,0927	0,0943	0,0959	0,0972	0,0986	0,0998	0,1010	0,1020
13	0,0706	0,0733	0,0759	0,0782	0,0804	0,0824	0,0842	0,0860	0,0876	0,0892	0,0906	0,0919	0,0932
14	0,0592	0,0622	0,0651	0,0677	0,0701	0,0724	0,0745	0,0765	0,0783	0,0801	0,0817	0,0832	0,0846
15	0,0481	0,0515	0,0546	0,0575	0,0602	0,0628	0,0651	0,0673	0,0694	0,0713	0,0731	0,0748	0,0764
16	0,0372	0,0409	0,0444	0,0476	0,0506	0,0534	0,0560	0,0584	0,0607	0,0628	0,0648	0,0667	0,0685
17	0,0264	0,0305	0,0343	0,0379	0,0411	0,0442	0,0471	0,0497	0,0522	0,0546	0,0568	0,0588	0,0608
18	0,0158	0,0203	0,0244	0,0283	0,0318	0,0352	0,0383	0,0412	0,0439	0,0465	0,0489	0,0511	0,0532
19	0,0053	0,0101	0,0146	0,0188	0,0227	0,0263	0,0296	0,0328	0,0357	0,0385	0,0411	0,0436	0,0459
20		0,0000	0,0049	0,0094	0,0136	0,0175	0,0211	0,0245	0,0277	0,0307	0,0335	0,0361	0,0386
21				0,0000	0,0045	0,0087	0,0126	0,0163	0,0197	0,0229	0,0259	0,0288	0,0314
22						0,0000	0,0042	0,0081	0,0118	0,0153	0,0185	0,0215	0,0244
23								0,0000	0,0039	0,0076	0,0111	0,0143	0,0174
24										0,0000	0,0037	0,0071	0,0104
25												0,0000	0,0350

Fonte: Teste de Shapiro-Wilk - Inferência | Portal Action. (n.d.). Retirado em 6 de Setembro de 2017, de http://www.portalaction.com.br/inferencia/64-teste-de-shapiro-wilk

Anexo 2.7.

Valores críticos de $d_{\alpha,n}$ de Kolmogorov-Smirnov

Teste unilateral						Teste unilateral					
$p =$	*0.90*	*0.95*	*0.975*	*0.99*	*0.995*	$p =$	*0.90*	*0.95*	*0.975*	*0.99*	*0.995*
Teste bilateral						**Teste bilateral**					
$p =$	*0.80*	*0.90*	*0.95*	*0.98*	*0.99*	$p =$	*0.80*	*0.90*	*0.95*	*0.98*	*0.99*
$n = 1$	.900	.950	.975	.990	.995	$n = 21$	.226	.259	.287	.321	.344
2	.684	.776	.842	.900	.929	**22**	.221	.253	.281	.314	.337
3	.565	.636	.708	.785	.829	**23**	.216	.247	.275	.307	.330
4	.493	.565	.624	.689	.734	**24**	.212	.242	.269	.301	.323
5	.447	.509	.563	.627	.669	**25**	.208	.238	.264	.295	.317
6	.410	.468	.519	.577	.617	**26**	.204	.233	.259	.290	.311
7	.381	.436	.483	.538	.576	**27**	.200	.229	.254	.284	.305
8	.358	.410	.454	.507	.542	**28**	.197	.225	.250	.279	.300
9	.339	.387	.430	.480	.513	**29**	.193	.221	.246	.275	.295
10	.323	.369	.409	.457	.489	**30**	.190	.218	.242	.270	.290
11	.308	.352	.391	.437	.468	**31**	.187	.214	.238	.266	.285
12	.296	.338	.375	.419	.449	**32**	.184	.211	.234	.262	.281
13	.285	.325	.361	.404	.432	**33**	.182	.208	.231	.258	.277
14	.275	.314	.349	.390	.418	**34**	.179	.205	.227	.254	.273
15	.266	.304	.338	.377	.404	**35**	.177	.202	.224	.251	.269
16	.258	.295	.327	.366	.392	**36**	.174	.199	.221	.247	.265
17	.250	.286	.318	.355	.381	**37**	.172	.196	.218	.244	.262
18	.244	.279	.309	.346	.371	**38**	.170	.194	.215	.241	.258
19	.237	.271	.301	.337	.361	**39**	.168	.191	.213	.238	.255
20	.232	.265	.294	.329	.352	**40**	.165	.189	.210	.235	.252
					Aproximação para $n > 40$		$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Fonte: Fonseca, J. (2001). *Estatística Matemática*. (Edições Sílaba, Ed.) (1a Edição). Lisboa.

Anexo 2.8.

Quantis da estatística de Lilliefors para a distribuição Normal

$p =$	*0.80*	*0.85*	*0.90*	*0.95*	*0.99*
$n = 4$	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231
25	.142	.147	.158	.173	.200
30	.131	.136	.144	.161	.187
> 30	$\frac{.736}{\sqrt{n}}$	$\frac{.768}{\sqrt{n}}$	$\frac{.805}{\sqrt{n}}$	$\frac{.886}{\sqrt{n}}$	$\frac{1.031}{\sqrt{n}}$

Fonte: Fonseca, J. (2001). *Estatística Matemática*. (Edições Sílaba, Ed.) (1a Edição). Lisboa.

Anexo 2.9.

Tabela de valores críticos de Durbin-Watson $\alpha = 0,05$

	k'*=1		k'=2		k'=3		k'=4		k'=5		k'=6		k'=7		k'=8		k'=9		k'=10	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.610	1.400	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.700	1.356	0.467	1.896	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.763	1.332	0.559	1.777	0.367	2.287	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645	0.203	3.004	-----	-----	-----	-----	-----	-----	-----	-----
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506	0.268	2.832	0.171	3.149	-----	-----	-----	-----	-----	-----
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390	0.328	2.692	0.230	2.985	0.147	3.266	-----	-----	-----	-----
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	-----	-----
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.471	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.258	0.502	2.461	0.407	2.668	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.691	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.731	2.124	0.637	2.290	0.546	2.461	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.735
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.750	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.013	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.959	0.874	2.071	0.798	2.188	0.723	2.309	0.649	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.681	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.796	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.079	1.891	1.015	1.978	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.876	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.197
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.838	1.369	1.874	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.747	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.846	1.608	1.862	1.593	1.877
200	1.758	1.779	1.748	1.789	1.738	1.799	1.728	1.809	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

*k' is the number of regressors excluding the intercept

Fonte: https://www3.nd.edu/~wevans1/econ30331/Durbin_Watson_tables.pdf. Retirado em 6 de Setembro de 2017.

Anexo 2.10.

Tabela valores críticos da Distribuição t-Studentized Range

$\alpha = .05$

ν \ k	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07	50.59	51.96	53.20	54.33	55.36	56.32	57.22	58.04	58.83	59.56
2	6.085	8.331	9.798	10.88	11.74	12.44	13.03	13.54	13.99	14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.37	16.57	16.77
3	4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462	9.717	9.946	10.15	10.35	10.53	10.69	10.84	10.98	11.11	11.24
4	3.927	5.040	5.757	6.287	6.707	7.053	7.347	7.602	7.826	8.027	8.208	8.373	8.525	8.664	8.794	8.914	9.028	9.134	9.233
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.802	6.995	7.168	7.324	7.466	7.596	7.717	7.828	7.932	8.030	8.122	8.208
6	3.461	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493	6.649	6.789	6.917	7.034	7.143	7.244	7.338	7.426	7.508	7.587
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.998	6.158	6.302	6.431	6.550	6.658	6.759	6.852	6.939	7.020	7.097	7.170
8	3.261	4.041	4.529	4.886	5.167	5.399	5.597	5.767	5.918	6.054	6.175	6.287	6.389	6.483	6.571	6.653	6.729	6.802	6.870
9	3.199	3.949	4.415	4.756	5.024	5.244	5.432	5.595	5.739	5.867	5.983	6.089	6.186	6.276	6.359	6.437	6.510	6.579	6.644
10	3.151	3.877	4.327	4.654	4.912	5.124	5.305	5.461	5.599	5.722	5.833	5.935	6.028	6.114	6.194	6.269	6.339	6.405	6.467
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.487	5.605	5.713	5.811	5.901	5.984	6.062	6.134	6.202	6.265	6.326
12	3.082	3.773	4.199	4.508	4.751	4.950	5.119	5.265	5.395	5.511	5.615	5.710	5.798	5.878	5.953	6.023	6.089	6.151	6.209
13	3.055	3.735	4.151	4.453	4.690	4.885	5.049	5.192	5.318	5.431	5.533	5.625	5.711	5.789	5.862	5.931	5.995	6.055	6.112
14	3.033	3.702	4.111	4.407	4.639	4.829	4.990	5.131	5.254	5.364	5.463	5.554	5.637	5.714	5.786	5.852	5.915	5.974	6.029
15	3.014	3.674	4.076	4.367	4.595	4.782	4.940	5.077	5.198	5.306	5.404	5.493	5.574	5.649	5.720	5.785	5.846	5.904	5.958
16	2.998	3.649	4.046	4.333	4.557	4.741	4.897	5.031	5.150	5.256	5.352	5.439	5.520	5.593	5.662	5.727	5.786	5.843	5.897
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108	5.212	5.307	5.392	5.471	5.544	5.612	5.675	5.734	5.790	5.842
18	2.971	3.609	3.997	4.277	4.495	4.673	4.824	4.956	5.071	5.174	5.267	5.352	5.429	5.501	5.568	5.630	5.688	5.743	5.794
19	2.960	3.593	3.977	4.253	4.469	4.645	4.794	4.924	5.038	5.140	5.231	5.315	5.391	5.462	5.528	5.589	5.647	5.701	5.752
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.896	5.008	5.108	5.199	5.282	5.357	5.427	5.493	5.553	5.610	5.663	5.714
24	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915	5.012	5.099	5.179	5.251	5.319	5.381	5.439	5.494	5.545	5.594
30	2.888	3.486	3.845	4.102	4.302	4.464	4.602	4.720	4.824	4.917	5.001	5.077	5.147	5.211	5.271	5.327	5.379	5.429	5.475
40	2.858	3.442	3.791	4.039	4.232	4.389	4.521	4.635	4.735	4.824	4.904	4.977	5.044	5.106	5.163	5.216	5.266	5.313	5.358
60	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646	4.732	4.808	4.878	4.942	5.001	5.056	5.107	5.154	5.199	5.241
120	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560	4.641	4.714	4.781	4.842	4.898	4.950	4.998	5.044	5.086	5.126
∞	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474	4.552	4.622	4.685	4.743	4.796	4.845	4.891	4.934	4.974	5.012

Fonte: Reis, E., Melo, P., Andrade, R., & Calapez, T. (2016). *Estatística Aplicada 2*. (Edições Sílabo, Ed.) (5a Edição). Lisboa

Anexo 2.11.

Tabela de quantis da estatística de Kruskal-Wallis para pequenas amostras

dimensão das amostras	$W_{0,90}$	$W_{0,95}$	$W_{0,99}$
2, 2, 2	3,7143	4,5714	4,5714
3, 2, 1	3 8571	4,2857	4,2857
3, 2, 2	4,4643	4,5000	5,3571
3, 3, 1	4,0000	4,5714	5,1429
3, 3, 2	4,2500	5,1389	6,2500
3, 3, 3	4,6000	5,0667	6,4889
4, 2, 1	4,0179	4,8214	4,8214
4, 2, 2	4,1667	5,1250	6,0000
4, 3, 1	3,8889	5,0000	5,8333
4, 3, 2	4,4444	5,4000	6,3000
4, 3, 3	4,7000	5,7273	6,7091
4, 4, 1	4,0667	4,8667	6,1667
4, 4, 2	4,4455	5,2364	6,8727
4, 4, 3	4,773	5,5758	7,1364
4, 4, 4	4,5000	5,6538	7,5385
5, 2, 1	4,0500	4,4500	5,2500
5, 2, 2	4,2933	5,0400	6,1333
5, 3, 1	3,8400	4,8711	6,4000
5, 3, 2	4,4946	5,1055	6,8218
5, 3, 3	4,4121	5,5152	6,9818
5, 4, 1	3,9600	4,8600	6,8400
5, 4, 2	4,5182	5,2682	7,1182
5, 4, 3	4,5231	5,6308	7,3949
5, 4, 4	4,6187	5,6176	7,7440
5, 5, 1	4,0364	4,9091	6,8364
5, 5, 2	4,5077	5,2462	7,2692
5, 5, 3	4,5363	5,6264	7,5429
5, 5, 4	4,5200	5,6429	7,7914
5, 5, 5	4,5000	5,6600	7,9800

Fonte: Reis, E., Melo, P., Andrade, R., & Calapez, T. (2016). *Estatística Aplicada 2*. (Edições Sílabo, Ed.) (5a Edição). Lisboa

Bibliografia

8 - Modelos Log Lineares para Tabelas de Contingência - Tabela Cruzada | Portal Action. (n.d.). Retrieved November 22, 2017, from http://www.portalaction.com.br/tabela-de-contingencia/modelos-log-lineares-para-tabelas-de-contingencia

Análise de Resíduos - Tabela Cruzada | Portal Action. (n.d.). Retrieved August 7, 2017, from http://www.portalaction.com.br/tabela-de-contingencia/analise-de-residuos

Eyduran, E., & Unit, B. G. (2005). Comparison of Chi-Square and Likelihood Ratio Chi-Square Tests: Power of Test 1, *1*(2), 242–244.

Fonseca, J. (2001). *Estatística Matemática*. (Edições Sílaba, Ed.) (1ª Edição). Lisboa.

Haberman, S. J. (1973). The Analysis of Residuals in Cross-Classified Tables. *Biometrics*, *29*(1), 205. https://doi.org/10.2307/2529686

Hall, A., Neves, C., & Pereira, A. (2011). *Grande Maratona de Estatística no SPSS*. (Escolar Editora, Ed.). Lisboa.

Hothorn, T., & Everitt, B. S. (2014). *A Handbook of Statistical Analyses Using R*. (CRC Press, Ed.) (third).

Howell, D. C. (2000). Chi-Square Test - Analysis of Contingency Tables, 1–4.

Leal, M. M. (1997). *Modelos Log-lineares em Tabelas de Contingência*. Lisboa.

Magalhães, F. M. de, Oliveira, C. T. de, & Silva, E. sá da. (2017). *Estatística Descritiva Aplicada à Gestão - Uma Análise Exploratória dos Dados*. (Vida Económica - Editorial, Ed.). Porto.

Matemática elementar/Estatística - Wikilivros. (n.d.). Retrieved November 21, 2017, from https://pt.wikibooks.org/wiki/Matemática_elementar/Estatística

McDonald, J. (n.d.). G–test of goodness-of-fit - Handbook of Biological Statistics. Retrieved November 22, 2017, from http://www.biostathandbook.com/gtestgof.html#chivsg

Mello, F. M. de. (2014). *Dicionário de estatística*. (L. Edições Sílabo, Ed.) (1ª edição). Lisboa.

Murteira, B., & Antunes, M. (2012). *Probabilidade e Estatística*. (Escolar Editora, Ed.) (Volume II). Lisboa.

Murteira, B., Ribeiro, C. S., Silva, J. A. e, Pimenta, C., & Pimenta, F. (n.d.). *Introdução à Estatística*.

Murteira, B., Ribeiro, C., Silva, J. e, & Pimenta, C. (2014). *Introdução à estatística*. Retrieved from http://pascal.iseg.utl.pt/~rui/Teaching/EI/stat1/errata.pdf

Pedrosa, A. C., & Gama, S. M. A. (2016). *Introdução computacional à probabilidade e estatistica com excel*. (Porto Editora, Ed.) (3ª Edição). Porto.

Pontos Influentes - Análise de Regressão | Portal Action. (n.d.). Retrieved August 16, 2017, from http://www.portalaction.com.br/analise-de-regressao/343-pontos-influentes

Reis, E. (2009). *Estatística Descritiva*. (Edições Sílaba, Ed.) (7ª Edição). Lisboa.

Reis, E., Melo, P., Andrade, R., & Calapez, T. (2016). *Estatística Aplicada 2*. (Edições Sílabo, Ed.) (5ª Edição). Lisboa.